RELAATIOTIETOKANTOJEN PITKÄAIKAISSÄILYTYS XML-KONVERSION AVULLA

SYSTEMAATTINEN KIRJALLISUUSKATSAUS

Heidi Eriksson

Tampereen yliopisto
Viestintätieteiden tiedekunta
Informaatiotutkimus ja interaktiivinen media
Pro gradu -tutkielma
Syyskuu 2017

TAMPEREEN YLIOPISTO, Viestintätieteiden tiedekunta
Informaatiotutkimus ja interaktiivinen media
ERIKSSON, HEIDI: Relaatiotietokantojen pitkäaikaissäilytys XML-konversion avulla. Systemaattinen kirjallisuuskatsaus.
Pro gradu -tutkielma, 58 s.
Syyskuu 2017

Tässä tutkielmassa tarkasteltiin relaatiotietokantojen pitkäaikaissäilytysmenetelmiä informaatiotutkimuksen näkökulmasta. Tutkielman painopisteenä oli XML-konversioon pohjautuvien pitkäaikaissäilytysmenetelmien käyttö relaatiotietokantojen pitkäaikaissäilytyksessä. XML-kielestä on muodostunut standardi väline digitaalisen tiedon kuvaamiseen, säilyttämiseen ja tiedon vaihtoon eri järjestelmien välillä. Tutkielman tavoitteena oli muodostaa systemaattisen kirjallisuuskatsauksen avulla kattava kuva siitä, millaisia menetelmiä relaatiotietokantojen XML-konversion toteuttamiseksi on pitkäaikaissäilytyksen näkökulmasta kehitetty, ja vertailla eri menetelmien soveltuvuutta relaatiotietokantojen pitkäaikaissäilytykseen. Vertailu perustui tutkielmassa määriteltyihin keskeisiin ominaisuuksiin, jotka käsittivät tietokannan tietosisällön ja relaatiorakenteen lisäksi tietokannan käyttäytymiseen liittyvät ominaisuudet. Tutkielmassa tarkasteltiin lisäksi sitä, millaisin eri tavoin säilytettyjen tietokantojen haku- ja prosessointimahdollisuudet oli eri menetelmissä toteutettu.

Tutkimuksessa tunnistettiin viisi erilaista XML-konversioon pohjautuvaa relaatiotietokantojen pitkäaikaissäilytysmenetelmää, joista kaksi oli prototyyppiasteella. Vertailu paljasti merkittäviä eroja säilytysmenetelmien toteutuksessa. Puutteita havaittiin erityisesti tietokannan käyttäytymiseen liittyvien ominaisuuksien säilyttämisen suhteen. Tutkimuksesta selvisi lisäksi, että säilytettävien tietokantojen haku- ja prosessointimahdollisuudet voidaan toteuttaa hyvin erilaisin tavoin riippuen säilytettävien tietokantojen käyttötarkoituksesta. Erityisen lupaavalta strategialta vaikuttaa tietokannan muuntaminen relaatiomallista yksinkertaisempaan tietomalliin säilytystä ja myöhempää tarkastelua varten tietovarastoteknologian avulla.

Tutkimus osoitti, että relaatiotietokantojen pitkäaikaissäilytyksen tutkimus on vielä varhaisessa vaiheessa. Tarjolla on hyvin vähän erilaisia menetelmiä ja työkaluja relaatiotietokantojen säilyttämiseksi. Aihepiirin tutkimus on keskittynyt säilytysformaattien tekniseen toteutukseen, ja liian vähän painoarvoa on annettu sille, kuinka pääsy säilytettävien tietokantojen tietosisältöön olisi toteutettavissa helppokäyttöisellä ja käyttäjäystävällisellä tavalla. Tulevaisuuden säilytysmenetelmiä suunniteltaessa on suositeltavaa huomioida tietokantojen käyttäytymiseen liittyvien ominaisuuksien säilyminen, jotta säilytettävien tietokantojen toiminnallisuutta ja käyttäytymistä kyetään tulkitsemaan tulevaisuudessakin. Jatkotutkimusta tarvitaan lisäksi tietovarastoteknologian hyödyntämisestä relaatiotietokantojen pitkäaikaissäilytyksessä, sekä pitkittäistutkimuksen mahdollistavista säilytysmenetelmistä ja hakutyökaluista.

Avainsanat: digitaalinen pitkäaikaissäilytys, relaatiotietokanta, systemaattinen kirjallisuuskatsaus, tietokanta, XML-konversio

Sisällysluettelo

1 JOHDANTO

Tietoteknologian voimakas kasvu viimeisten vuosikymmenien aikana on johtanut tilanteeseen, jossa suuri osa ihmiskunnan tietämyksestä on tallennettuna digitaaliseen muotoon. Tämä 1900-luvun loppupuolella alkanut kehitys on tuonut mukanaan tilanteen, jossa valtavat määrät tietoa on tallennettuna digitaalisiin järjestelmiin lukuisille erilaisille alustoille ja ohjelmistoille lukuisissa erilaisissa tiedostomuodoissa. Nämä alustat muuttuvat ja kehittyvät jatkuvasti ja nopeasti, mikä on synnyttänyt uudenlaisia haasteita myös digitaalisen tiedon säilyttämiselle. Digitaalisen tiedon säilymistä uhkaavat monet tekijät, kuten säilytysmedian rappeutuminen, ohjelmistojen ja tiedostomuotojen vanhentuminen, sekä olennaisten metatietojen puuttuminen, jolloin tieto-objektin merkitys voidaan menettää ja sen todistusarvo vaarantua.

Suuri osa digitaalisiin järjestelmiin tallennetusta tiedosta sijaitsee *tietokannoissa* (engl. *database*). Tietokannat on suunniteltu tiedon tallentamiseen, organisointiin ja hakemiseen, ja ne ovat perustavanlaatuinen osa lähes kaikkia tietojärjestelmiämme. Tietokantoihin tallennettu tieto on usein korvaamatonta, ja sitä voi olla mahdoton tuottaa uudelleen, jos se menetetään. Tämän vuoksi on elintärkeää, että tietokantoihin tallennettu tieto säilytetään asianmukaisissa säilytysjärjestelmissä. Suurin osa digitaalisen säilytyksen tutkimuksesta on kuitenkin keskittynyt tavanomaisten tiedostomuotojen, kuten tekstitiedostojen, kuvien ja äänitiedostojen pitkäaikaissäilytykseen. Tietokantojen pitkäaikaissäilytystä on tutkittu hyvin vähän siitä huolimatta, että ne ovat keskeisessä roolissa niin julkishallinnon kuin liiketoiminnan organisaatioidenkin tehtävien hoidossa ja asiakirjanhallinnassa, sekä yliopistojen ja tutkimuslaitosten tuottaman tieteellisen datan hallinnoimisessa.

Useimmiten tietokantojen, kuten muunkin digitaalisen datan, arkistointia motivoi lainsäädännöstä tai liiketoiminnan vaatimuksista kumpuava tarve säilyttää dataa todisteena eli *evidenssinä* (engl. *evidence*) organisaation toiminnasta. Suomessa Kansallisarkisto määrää, mitkä julkishallinnon viranomaisten asiakirjat säilytetään pysyvästi (Arkistolaki 1994/831 § 8). Tietokantojen arkistoinnista saadaan myös muita hyötyjä, joista tietokantajärjestelmän nopeutuminen ja tehostuminen, sekä siitä seuraavat kustannussäästöt ovat vain osa. Tieteellisten data-arkistojen tietokannat toimivat ainutlaatuisena tutkimusaineistona, ja arkistoihin tallennetut otokset julkishallinnon tietokannoista voivat tarjota

arvokasta tietoa kulttuurihistoriamme kannalta tärkeiden tapahtumien kehittymisestä ja muuttumisesta pitkällä aikavälillä. (ks. mm. Müller 2009.)

Viimeisten kahden vuosikymmenen aikana niin kutsuttujen *relaatiotietokantojen* (engl. *relational database*) käyttö on yleistynyt voimakkaasti yritysten ja julkishallinnon organisaatioiden tietojärjestelmissä. Tämä asettaa haasteita kaikille niille organisaatioille, jotka ovat vastuussa asiakirjallisen tiedon kokoamisesta ja säilyttämisestä. Keskiössä eivät ole vain kansallisarkistot, -kirjastot, ja tieteelliset data-arkistot, vaan kaikki ne organisaatiot ja liiketoimintayritykset, joiden tulee täyttää lailliset velvoitteensa. Relaatiotietokantajärjestelmistä on kehittynyt valtavan monimutkaisia järjestelmiä, minkä johdosta datan käsittely alkuperäisen tietokantajärjestelmän ulkopuolella ei ole mahdollista ilman automatisoituja työkaluja. Kansallisarkistoille ja muille keskusarkistoille tulee dataa eri tietokantajärjestelmistä, mikä luo paineita yhtenäisten käytäntöjen ja standardien luomiseksi, jotta eri lähteistä tuleva data saadaan otettua talteen, ja säilytettyä käytettävänä ja saavutettavana vuosikymmenien ajan. (ks. mm. Heuscher ym. 2004, 1–2.)

Tänä päivänä yleisesti käytetty menetelmä relaatiotietokantojen säilyttämiseksi on tietokannan *konversio* (engl. *conversion*) eli muuntaminen pitkäaikaissäilytyksen kannalta sopivaan tiedostomuotoon. *XML (eXtensive Markup Language)* on avoin, neutraali tiedostomuoto, jota on käytetty tiedon kuvaamiseen ja tiedonvaihtoon 1990-luvun lopulta lähtien. Se on yleistynyt myös digitaalisessa pitkäaikaissäilytyksessä tiedon *säilytysmuotona* (engl. *preservation format*), sekä *metatietojen* (engl. *metadata*) esitysmuotona.

Tässä tutkielmassa keskitytään tarkastelemaan XML:n käyttöä relaatiotietokantojen pitkäaikaissäilytyksessä. Tutkielman tarkoituksena on kartoittaa XML-konversioon pohjautuvien relaatiotietokantojen pitkäaikaissäilytysmenetelmien kirjo systemaattisen kirjallisuuskatsauksen avulla. Tavoitteena on tuottaa kirjallisuuden pohjalta kattava kuva siitä, millaisia XML-konversioon pohjautuvia relaatiotietokantojen säilytysratkaisuja on tähän mennessä kehitetty, sekä vertailla eri menetelmien soveltuvuutta tarkoitukseen. Tutkielmassa tarkastellaan lisäksi sitä, kuinka pääsy arkistoidun tietokannan tietosisältöön on eri säilytysmenetelmissä toteutettu. Pyrkimyksenä on tuoda aihepiiriä alan tutkijoiden ja ammattilaisten tietouteen; tunnistaa alueita, joissa on tarvetta lisätutkimukselle; sekä tuottaa tarpeellista ja ajankohtaista taustatietoa, josta on hyötyä yritysten ja julkishallinnon organisaatioiden relaatiotietokantojen säilytysstrategioiden kehittämisessä.

2 TIETOKANTOJEN PITKÄAIKAISSÄILYTYS

Digitaalisen pitkäaikaissäilytyksen (engl. *digital preservation*) tavoitteena on suojella kulttuurihistoriallisesti arvokkaaksi katsottua tietosisältöä – olipa se sitten alkuperältään digitaalisessa muodossa, tai digitaaliseen muotoon siirrettyä – ja tarjota siihen pääsy nykyisille ja tuleville sukupolville (ks. mm. Conway 2010). Tässä tutkielmassa digitaalisella pitkäaikaissäilytyksellä tarkoitetaan niitä toimintoja ja prosesseja, jotka tähtäävät informaation sekä muun kulttuuriperinnön säilyttämiseen digitaalisessa muodossa pitkäaikaisesti, ja takaavat pääsyn kyseiseen tietoon ihmisille ymmärrettävässä muodossa. *Tietokantojen pitkäaikaissäilytys* (engl. *database preservation*) on digitaalisen pitkäaikaissäilytyksen osa-alue, jonka tarkoituksena on varmistaa tietokantaan tallennetun datan ja tietokannan rakenteen säilyminen, sekä tietokannan käytettävyys ja saavutettavuus pitkällä aikavälillä.

Ennen kuin perehdymme relaatiotietokantojen pitkäaikaissäilytykseen lähemmin, on tarkasteltava relaatiotietokantojen ominaispiirteitä, jotta voimme ymmärtää, kuinka nämä ominaispiirteet vaikuttavat säilytysratkaisun valintaan.

2.1 Johdatus relaatiotietokantoihin

Tietokanta on tiettyä tarkoitusta varten kehitetty kokoelma loogisesti yhteenkuuluvaa tietoa, jota useat eri käyttäjät voivat käyttää samanaikaisesti. Tietokannat koostuvat kaksiulotteisista taulukoista eli *tauluista* (engl. *table*), joiden avulla tietoa jäsennetään. Taulun sisältämiä rivejä kutsutaan *tietueiksi* (engl. *tuple, record*) ja sarakkeita *attribuuteiksi* (engl. *attribute*) tai *kentiksi* (engl. *field*). Jokaiselle taulun attribuutille määritellään *tietotyyppi* (engl. *data type*), joka määrää, millaisia *arvoja* (engl. *value*) attribuutti voi saada. Tietokantojen rakennetta puolestaan kutsutaan *skeemaksi* (engl. *schema*). (Connolly & Begg 2015, 63–68.)

Yksinkertaisimmillaan tietokanta voi koostua yhdestä taulusta, jossa on muutamia rivejä tietoa. Yleensä tietokanta kuitenkin koostuu kymmenistä tauluista, ja yksittäisissä tauluissa voi olla jopa miljoonia rivejä tietoa. Tietokannat voivat sisältää numeroiden ja tekstin lisäksi monimutkaisempaa dataa, kuten kokonaisia tiedostoja. Taulukossa 1 on ote kuvitteellisen tietokannan *Työntekijä*-taulun sisällöstä. Taululla on viisi attribuuttia

– *Id, Etunimi, Sukunimi, Syntymaaika, Osasto* – ja se sisältää neljä tietuetta, jotka kukin saavat viisi arvoa, yhden kutakin attribuuttia kohden.

Id	Etunimi	Sukunimi	Syntymaaika	Osasto
1	Miina	Matikainen	6.6.1956	1
2	Tuukka	Nousiainen	2.1.1965	3
3	Risto	Mäenpää	3.9.1979	2
4	Mathilda	Hansson	14.6.1974	2

Taulukko 1: Kuvitteellisen Työntekijä-taulun sisältöä.

Tietokannan rakenne perustuu yleensä johonkin matemaattiseen malliin, joka määrää sen, millä tavalla tietokantaan voidaan tallentaa tietoa. Tällä hetkellä yleisimmin käytetty tietokantatyyppi ovat relaatiotietokannat (ks. Dell Software 2015). Relaatiotietokantojen historia ulottuu 1970-luvulle, jolloin Edgar Frank Codd kehitteli tietokantojen relaatiomallin osana IBM:n tutkimusta (Codd 1970). Relaatiomalli pysyi pitkään teoreettisena, kunnes tietotekniikan kehitys mahdollisti kaupallisten tietokantaohjelmistojen yleistymisen 1980-luvulla

Relaatiotietokantojen ominaispiirteitä ovat tietokannan taulujen väliset yhteydet. Relaatiotietokannan tauluja kutsutaankin *relaatioiksi* (*engl. relation*) matemaattisen relaatiomallin mukaan. Relaatiotietokannassa tietoa muodostetaan yhdistelemällä eri taulujen tietueisiin tallennettua dataa. Taulun jokaisen tietueen tulee olla yksilöllinen. Tietueen yksilöllisyyden takaa niin kutsuttu *perusavain* (engl. *primary key*), jonka avulla taulun jokainen tietue voidaan tunnistaa yksikäsitteisesti. Esimerkiksi henkilötietokannassa perusavaimena voi toimia henkilötunnus, joka on jokaiselle henkilölle yksilöllinen. Perusavaimena voi toimia myös juokseva tunnusnumero. Perusavaimen arvo ei voi koskaan olla tyhjä; tätä kutsutaan *avaineheydeksi* (engl. *entity integrity*). (Connolly & Begg 2015, 152, 162.)

Perusavaimen lisäksi relaatiotietokannassa voi olla *viiteavaimia* (engl. *foreign key*), jotka kuvaavat yhteyksiä tietokannan taulujen välillä. Yhteys toisen taulun perusavaimeen osoitetaan viiteavaimella. Jokaista viittaavassa taulussa esiintyvää viiteavaimen arvoa tulee vastata sama perusavaimen arvo viitattavassa taulussa. Taulua, jolla on viittauksia toisiin tauluihin, ei voida poistaa. Tätä kutsutaan *viite-eheydeksi* (engl. *referential integ-*

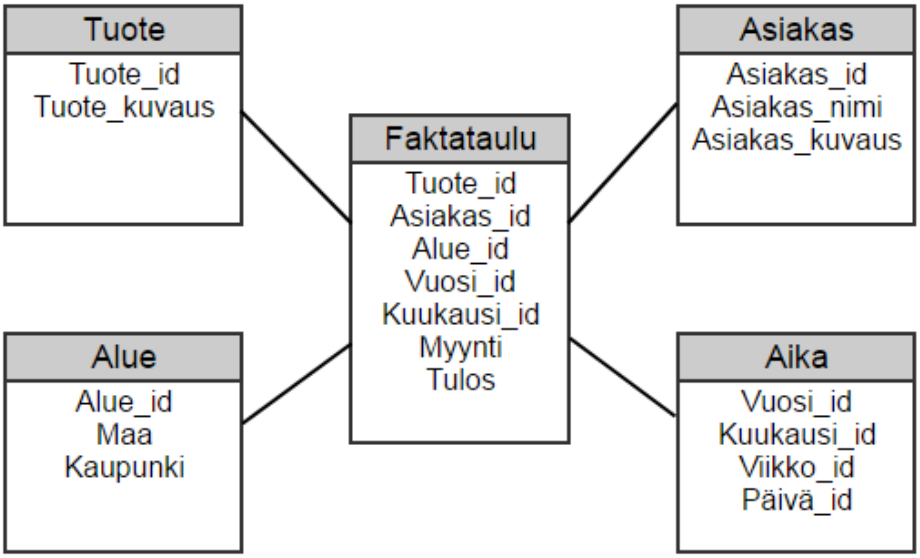
rity). (Connolly & Begg 2015, 152.) Tietokannan taulujen välisiä viittauksia on havainnollistettu kuviossa 1.

Kuvio 1: Projekti-taulun viiteavain Projektipaallikko viittaa Työntekijä-taulun perusavaimeen. Työntekijä-taulun viiteavain Osasto viittaa Osasto-taulun perusavaimeen. Perusavaimet on merkitty lihavoinnilla.

Relaatiotietokannoissa tietojen toistumista ja siitä aiheutuvia ongelmia tietokannan tietojen päivityksessä pyritään välttämään niin kutsutulla *normalisoinnilla* (engl. *normalization*). Normalisoinnissa tietokannan rakennetta yksinkertaistetaan siten, että kukin attribuutti sisältää vain yhden arvon, ja kukin attribuutin arvo sijaitsee vain yhdessä paikassa. Tavoitteena on pilkkoa tietokannan sisältämä tieto mahdollisimman pieniin palasiin, sekä vähentää saman tiedon toistumista eri tauluissa. Normalisoinnin avulla tietokantaa on helpompi ylläpitää; siitä saadaan selkeämpi, yhdenmukaisempi ja joustavampi. (Connolly & Begg 2015, 452.)

Joskus normalisoinnin purkaminen voi olla perusteltua. Näin toimitaan esimerkiksi *tietovarastojen* (engl. *data warehouse*) tapauksessa. Tietovarastolla tarkoitetaan eräänlaista analyysitietokantaa, johon ladataan sopivalla tavalla koostettua tietoa organisaation operatiivisista tietokannoista ja muista tietolähteistä. Tietovarastot on tarkoitettu erityisesti suurten tietomäärien tallennusta ja analysointia varten, ja ne eroavat tietomalliltaan relaatiotietokannoista. Tavoitteena on tuottaa tietokantojen sisällöstä yksinkertaisempi esitystapa, jota voidaan käyttää organisaatiossa päätöksenteon tukena. Tietovarastoissa tietoja ryhmitellään *ulottuvuuksiin* (engl. *dimension*) eri tekijöiden, kuten tuotteiden, asiakkaiden tai ajan, perusteella. Nämä ulottuvuudet kytkeytyvät niin kutsuttuun *faktatauluun* (engl. *fact table*), joka sisältää mitattavana olevan datan arvot. Tällaista moniulotteista tietomallia kutsutaan yleensä *tähtimalliksi* (engl. *star schema*) tähteä muistut-

tavan muotonsa vuoksi. Tähtimallia on havainnollistettu kuviossa 2. (Aldeias ym. 2011, 117.)



Kuvio 2: Faktataulu ja siihen kytkeytyvät ulottuvuudet.

Tietokannat vaativat *tietokannanhallintajärjestelmän* (engl. *database management system*) toimiakseen. Tietokannanhallintajärjestelmä on ohjelmisto, jonka avulla tietokantaa voidaan ylläpitää, hallita ja päivittää. Tietokannanhallintajärjestelmän avulla voidaan myös välittää tietokannan sisältämää tietoa muille sovelluksille. *Tietokantajärjestelmällä* (engl. *database system*) puolestaan tyypillisesti tarkoitetaan tietokannan, tietokannanhallintajärjestelmän, sekä tietokannan kanssa vuorovaikutuksessa olevien tietokantasovellusten muodostamaa kokonaisuutta. (Connolly & Begg 2015, 52.) Tunnetuimpia kaupallisia relaatiotietokannanhallintajärjestelmiä ovat Firebird, Microsoft Access, Microsoft SQL Server, MySQL, Oracle ja PostgreSQL.

Tietoa relaatiotietokannasta haetaan standardoidun *SQL-kyselykielen* (*Structured Query Language*) avulla. SQL-kyselykieli kehitettiin IBM:llä 1970-luvulla relaatiomalliin pohjautuvan IBM System R -nimisen tietokantajärjestelmän prototyypin hallinnoimiseksi (Elmasri & Navathe 1994, 185). SQL-kyselykielellä on keskeinen rooli relaatiotietokannan hallinnassa. SQL-kielen avulla voidaan tietokannan tauluja luoda, päivittää ja poistaa. Yksinkertainen SQL-kielinen lauseke *Tyontekija*-nimisen taulun luomiseksi on esitetty kuviossa 3. Tässä esimerkissä attribuutti *'Id'* on määritelty kokonaisluvuksi, ja attribuutit *'Etunimi'* ja *'Sukunimi'* merkkijonoiksi, joiden pituus voi olla korkeintaan 100

merkkiä. ’*Id*’ määritellään taulun perusavaimeksi. ’*Osasto’* määritellään viiteavaimeksi, joka viittaa *Osasto*-taulun *’Id’*-kenttään.

```
CREATE TABLE Tyontekija (
      Id integer NOT NULL,
      Etunimi varchar(100) NOT NULL,
      Sukunimi varchar(100) NOT NULL,
           PRIMARY KEY (Id),
      FOREIGN KEY (Osasto) REFERENCES Osasto(Id)
);
```

Kuvio 3: SQL-kielinen lauseke.

SQL mahdollistaa myös erilaisten *näkymien* (engl. *view*) luomisen tietokannan sisältämästä tiedosta. Näkymät ovat virtuaalisia tauluja, jotka on koottu yhteen SQL-kyselyiden avulla tietokannan muiden taulujen tai näkymien pohjalta. Käyttäjän kannalta tietokantanäkymä toimii, kuten mikä tahansa tietokannan perustaulu. Erona on, että tietokantanäkymiä ei välttämättä pystytä päivittämään. (Pulkkinen 1994, 55.)

SQL sai ANSI-standardin vuonna 1986 ja ISO-standardin vuonna 1987 (ISO 9075:1987). Tämän jälkeen standardiin on kohdistunut useita uudistuksia, joista viimeisin vuonna 2016 (ISO/IEC 9075-1:2016). Standardoinnista huolimatta tietokantajärjestelmien toteutukset eri valmistajien välillä eroavat toisistaan suuresti, sillä SQL-standardi jättää valmistajalle paljon liikkumavaraa. Valmistajat voivat muun muassa määritellä omia tietotyyppejään, aliohjelmiaan ja operaattoreitaan. Eri valmistajat ovat toteuttaneet omia versioitaan tietokantaan tallennettavista *herättimistä* (engl. *triggers*), *funktioista* (engl. *function*) ja *tallennetuista proseduureista* (engl. *stored procedure*). Herättimet ovat toimintoja, jotka suoritetaan automaattisesti, kun tietylle taululle suoritetaan jokin tietty SQL-operaatio. Herättimiä käytetään tyypillisesti tietokannan eheyden ylläpitämiseen lisäys-, poisto- ja päivitysoperaatioiden yhteydessä. Tallennetut proseduurit ja funktiot puolestaan ovat tietokantajärjestelmään tallennettuja aliohjelmia, joiden avulla pyritään tehostamaan SQL-kyselyitä ja nopeuttamaan kyselyiden suorittamiseen kuluvaa aikaa. Tallennettujen proseduurien kieli ja syntaksi riippuvat valmistajasta, eivätkä kaikki järjestelmät tue niitä lainkaan. (Heuscher ym. 2004, 3; Connolly & Begg 2015, 280–281.). Standardeista poikkeaminen hankaloittaa tietokannan muuntamista järjestelmästä toiseen – myös arkistoitavaan muotoon. Tietokantojen pitkäaikaissäilytyksessä

huomioitavia seikkoja käsitellään yksityiskohtaisemmin seuraavissa alaluvuissa. Aivan ensin tarkastellaan digitaalisen pitkäaikaissäilytyksen edellytyksiä yleisellä tasolla.

2.2 Digitaalisen pitkäaikaissäilytyksen edellytykset

Digitaalisen aineiston arkistoiminen eroaa perinteisestä arkistoinnista monella tavalla. Digitaalisella aineistolla tarkoitetaan tässä mitä tahansa binäärisessä muodossa tallennettua tietoa. Tämä kattaa niin tekstidokumentit, kuva- ja äänitiedostot, www-sivut, ohjelmistot kuin tietokannatkin. Paperilla oleva asiakirja on arkistoinnin kannalta helppohoitoinen. Paperi sijoitetaan fyysisesti sen lopulliseen säilytyspaikkaan tietyn järjestelmän mukaan luokiteltuna. Suotuisissa olosuhteissa tieto säilyy luettavana jopa vuosisatoja ilman erityisiä toimenpiteitä (Henttonen 1999, 23). Digitaalinen tieto on aina sidoksissa tiettyyn ohjelmisto- ja laiteympäristöön. Kun tarkoituksena on säilyttää tieto siten, että se on käytettävissä pitkiäkin ajanjaksoja, muodostuu avainongelmaksi se, että säilytysaika voi ylittää käytössä olevan tallennusmedian, laitteiston, ohjelmiston ja tiedostoformaattien eliniän (ks. mm. Lin ym. 2003, 117).

2.2.1 Aineiston fyysinen säilyminen

Fyysisen säilymisen varmistaminen on digitaalisen pitkäaikaissäilytyksen perusedellytyksiä (Henttonen 1999, 27). Kaikki fyysinen tallennusmedia on altista eroosiolle. Optinen ja magneettinen media, kuten CD- ja DVD-levyt sekä magneettinauhat, ovat alttiita naarmuuntumiselle sekä kosteuden ja lämpötilan vaihteluille. Kiintolevyissä on kuluvia osia ja Flash-muistin muistipaikat kuluvat käytössä. CD- ja DVD-levyjen keskimääräinen elinikä on vain 2-5 vuotta, Blu-Ray-levyjen 10-15 vuotta. Magneettinauhojen eliniäksi on arvioitu noin 10-30 vuotta. Verrattuna arkistointiin tarkoitettuun paperiin tai mikrofilmiin, jonka elinikä voi olla jopa 500 vuotta, tuntuvat nämä mediat varsin tilapäisiltä ratkaisuilta. (Hedstrom 1998, 197–198; Atos 2014, 8.) Yleensä digitaalisen tiedon pitkäaikaissäilytyksessä käytetään magneettisia ja optisia tallennusmedioita. Kiintolevyt eivät sovellu pitkäaikaiseen säilyttämiseen. Tallennusmedioiden rappeutumisen vuoksi suositellaan tiedon siirtämistä uudelle tallennusmedialle tietyin aikavälein, esimerkiksi joka viides vuosi. (Lybeck 2006, 127.) Tätä toimenpidettä kutsutaan tallennusmedian *virkistämiseksi* (engl. *refreshment*) tai *tuoreuttamiseksi*, ja se on muodostunut vallitsevaksi käytännöksi.

2.2.2 Aineiston luettavuus, käytettävyys ja ymmärrettävyys

Fyysisen säilymisen lisäksi on voitava varmistua aineiston *luettavuudesta*, *käytettävyydestä* ja *ymmärrettävyydestä* pitkällä aikavälillä (Henttonen 1999, 27–28). Suurin uhka aineiston luettavuudelle on laitteistojen, ohjelmistojen ja tiedostoformaattien vanheneminen. Uusia laitteisto- ja ohjelmistoversioita syntyy 3-5 vuoden välein, eivätkä uudet versiot ole aina yhteensopivia vanhojen kanssa. (Hedstrom 1998, 191.) Vielä 1990-luvun lopulla yleisesti käytössä olleiden 3½-tuuman levykkeiden lukemiseksi voi nykyään olla vaikea löytää laitteistoa. Vaikka sopiva levykeasema löytyisikin, ei ole takeita siitä, että levykkeeltä löytyvä tiedosto aukeaisi jollakin tämän päivän ohjelmalla. Tiedostoformaattien vanhenemiseen käytetään usein ratkaisuna tiedostojen konvertoimista eli muuntamista standardoituun muotoon (Hakala 2002, 24). Tiedosto konvertoidaan johonkin yleisesti käytettyyn, mielellään avoimeen standardiin, jonka oletetaan säilyvän yleisessä käytössä ja olennaisilta osiltaan muuttumattomana pitkään.

Ei riitä, että aineisto on luettavissa, vaan sen tulee olla käytettävissä ja muunnettavissa ihmisen ymmärtämään muotoon. Haluttu tieto on voitava paikantaa tietokannasta ja sitä on voitava käsitellä. Relaatiotietokanta koostuu tietyn mallin mukaan järjestetyn tietosisällön lisäksi tietokannanhallintajärjestelmästä, sekä yhdestä tai useammasta tietokantasovelluksesta, joiden avulla tietokannan sisältöä voidaan lisätä, poistaa, muokata ja, ennen kaikkea, havainnollistaa ihmisen ymmärtämään muotoon. Tieto on pilkottu pieniin, toisistaan riippumattomiin osiin, jotka sijaitsevat tietokannan eri tauluissa, ja vasta käyttöhetkellä kootaan yhteen ymmärrettäväksi kokonaisuudeksi (Pulkkinen 1994, 53). Tietokannasta voidaan kyllä ottaa kopio, mutta ilman alkuperäistä tietokantajärjestelmää ja tuotantoympäristöä on tietokannan palauttaminen käyttökelpoiseksi – ja sen sisällön tulkitseminen – erittäin vaikeaa (Hakala 2002, 22). Tulostettaessa tietokanta paperille tai tallennettaessa se peräkkäistiedostoksi menetetään kaikki tietokannan prosessointimahdollisuudet. Tietokannan sisältö on kyllä tallessa ja luettavissakin - mutta ei käytettävissä.

Käytettävyyden lisäksi on voitava varmistua aineiston ymmärrettävyydestä. Jos tietokannan sisältö ja käyttötarkoitus eivät ole tiedossa, jää tietokannan merkitys epäselväksi, eikä tietokannan sisältämää dataa voida tulkita. Aineiston ymmärrettävyys varmistetaan riittävällä ja huolellisella dokumentoinnilla. (Henttonen 1999, 27–28.)

2.2.3 Aineiston dokumentointi

Fyysisen säilymisen, käytettävyyden ja ymmärrettävyyden lisäksi tulee säilytysprosessissa huomioida aineiston *eheyden* (engl. *integrity*) rikkomattomuus. Eheydellä tarkoitetaan sitä, että digitaalinen asiakirja on aito, luotettava ja keskeisiltä ominaisuuksiltaan muuttumaton. Luotettavuudella tarkoitetaan tässä yhteydessä asiakirjan luotettavuutta todisteena arkistonmuodostajan toiminnasta. Eheyteen liittyy läheisesti *autenttisuuden* (engl. *authenticity*) eli alkuperäisyyden käsite: onko asiakirja sitä, mitä se sanoo olevansa? Onko asiakirjan lähettäjä se, joka tämä väittää olevansa? (ks. Lybeck 2006, 125–126.)

Digitaalisten järjestelmien luonteen vuoksi digitaalisen asiakirjan eheyden ja aitouden määrittäminen ei ole yksiselitteistä, sillä tieto on jatkuvasti muokattavissa, eikä muokkaamisesta välttämättä jää jälkeä itse tiedostoon. Digitaalisen asiakirjan arkistointiin sisältyy lisäksi lähes aina bittitason muutoksia, sillä tiedostoa joudutaan yleensä muokkaamaan, jotta se voidaan muuntaa säilytysjärjestelmän vaatimaan muotoon. Aina, kun asiakirjaan joudutaan tekemään muutoksia, vaarantuu sen aitous ja luotettavuus. (ks. Factor ym. 2009.)

Digitaalisen asiakirjan autenttisuuden ja luotettavuuden takaa ensisijaisesti asiakirjanhallintaprosessi, jossa asiakirjaa on säilytetty asianmukaisissa järjestelmissä. Asiakirjan konteksti sekä tiedostolle tehdyt muutokset tulee dokumentoida tarkoin aina asiakirjan luomishetkestä lähtien, jotta asiakirjan eheys ja koskemattomuus voidaan taata (ks. Factor ym. 2009). Asiakirjan eheys ja autenttisuus varmistetaan tavallisesti liittämällä asiakirjaan *metatietoa* (engl. *metadata*). Metatiedon avulla asiakirja voidaan sitoa aikaan, paikkaan, ja siihen toimenpiteeseen, johon se liittyy. Metatieto palvelee myös aineiston hakua järjestelmistä. Metatiedolla on digitaalisessa tiedonsäilytyksessä kriittinen rooli; ilman metatietoa ei aineiston alkuperää voida vahvistaa. Epätäydellinen tai puutteellinen metatieto hankaloittaa aineiston esittämistä alkuperäisessä muodossa. Kriittisen metatiedon puuttuminen voi pahimmassa tapauksessa tarkoittaa sitä, että dokumentti ei kelpaa todisteeksi toiminnasta, eikä siten voi toimia asiakirjana. (Lin ym. 2003, 120; Lybeck 2006, 73.)

Digitaalisen asiakirjan vaatimat metatiedot voidaan Dappertin ja Endersin (2010, 6) mukaan ryhmitellä neljään kategoriaan: *kuvaileva*, *rakenteellinen*, *tekninen* ja *hallinnolli-*

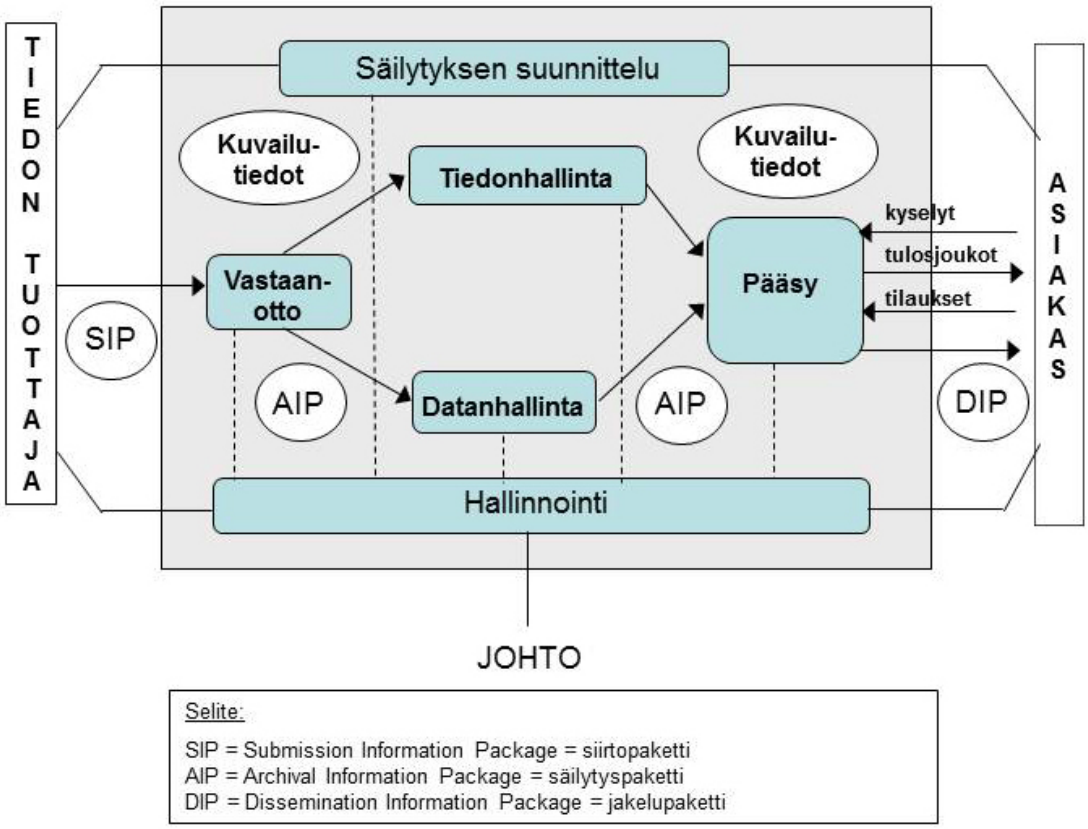
nen metatieto. Kuvaileva metatieto kuvaa asiakirjan sisällön: tekijän, otsikon, ja muita tietoja, jotka auttavat tiedon löytämisessä. Se kertoo, mikä asiakirjan tarkoitus oli, ja tallentaa historiallista kontekstia asiakirjan alkuperästä. Rakenteellinen metatieto kuvaa asiakirjan sisältämät fyysiset ja loogiset rakenteet, kuten mikä kuva on tallennettu millekin www-sivulle, tai mikä sivu seuraa toista elektronisessa kirjassa. Tekninen metatieto sisältää tiedostotyyppiin liittyvää tietoa: millä ohjelmistolla ja laitteistolla tiedoston voi avata ja esittää. Tämän lisäksi tekniseen metatietoon tallentuu jälki digitaalisen asiakirjan sisältämän tiedon muuttumattomuudesta ja autenttisuudesta. Tekniseen tietoon sisältyy myös asiakirjan sisältöön liittyvää teknistä tietoa, kuten tieto kuvan koosta tai äänitiedoston pituudesta. Hallinnollinen metatieto puolestaan säilyttää tietoa asiakirjan provenanssista: kuka asiakirjaa on ylläpitänyt, ja mitä arkistointitoimenpiteitä sille on tehty. Hallinnolliseen metatietoon sisältyy myös tiedot asiakirjan käyttöoikeuksista ja luvat toimenpiteistä, joita sille voidaan tehdä. (Dappert & Enders 2010, 6.)

Suomessa julkishallinnon asiakirjojen metatietovaatimukset on määritelty Kansallisarkiston SÄHKE2-määräyksessä sekä julkisen hallinnon tietohallinnon neuvottelukunta JUHTA:n julkaisemassa JHS 143 -suosituksessa (Arkistolaitos 2009; JUHTA 2004). Toisin kuin muille digitaalisille asiakirjoille, ei Suomen julkishallinnossa vielä ole metatietomääritystä rekisteritiedolle, johon myös tietokannat luetaan. JUHTA on käynnistänyt hankkeen rekisteritietojen metatietojen määrittelemiseksi. Hankkeen tavoitteena on luoda yhdenmukaiset metatietomääritykset julkishallinnon organisaatioiden rekisteritiedon kuvaamiselle JHS-suosituksen muodossa. (JUHTA 2017.)

2.3 OAIS-viitemalli

Digitaalisesta pitkäaikaissäilytyksestä puhuttaessa on hyvä esitellä digitaalisen säilytyksen *OAIS (Open Archival Information System)* -viitemalli, joka on yleistynyt laajaan käyttöön. Tämä CCSDS:n *(The Consultative Committee for Space Data Systems Standards)* kehittämä viitemalli määrittelee kaikki pitkäaikaissäilytysarkistoon eli PAS-arkistoon liittyvät keskeiset toiminnot, toimijat ja niiden väliset suhteet. Mallin mukaisesti *tuottaja* tuottaa informaatiota, jonka se välittää *siirtopakettina* (lyh. *SIP*) *hallinnoijalle*, eli arkistolle. Arkisto vastaanottaa siirtopaketin, joka muunnetaan *säilytyspaketiksi* (lyh. *AIP*). Säilytyspaketti muunnetaan edelleen jakelupaketiksi (lyh. *DIP*), joka tarjotaan *ku-*

luttajille. (Ferreira 2016, 7-8; Hakala 2014.) Viitemallin keskeiset toiminnalliset osat on kuvattu kuviossa 4.



Kuvio 4: PAS-arkiston toiminnalliset osat. (Hakala 2014.)

Viitemallin suomalainen versio, SFS 5972, julkaistiin vuonna 2009, ja se on päivitetty vuonna 2015 vastaamaan vuonna 2012 ilmestynyttä OAIS-viitemallia (Hakala 2014). OAIS-viitemalli on hyvä tuntea myös tietokantojen pitkäaikaissäilytyksestä puhuttaessa, sillä useimmat tietokantojen säilytysformaatit on suunniteltu toimimaan OAIS-mallin mukaisena siirtopakettina.

2.4 Keskeisten ominaisuuksien valinta

Kun tietokanta arkistoidaan, se siirretään pois alkuperäisestä ympäristöstään. Tietokannan sisällöstä, rakenteesta tai toiminnallisuudesta jää väistämättä joitakin ominaisuuksia puuttumaan, ellei koko tietokannanhallintajärjestelmää ja muuta käyttöympäristöä säilytetä. Tietokannan säilytystä suunniteltaessa on valittava ne tietokannan *keskeiset ominaisuudet* (engl. *significant properties, significant characteristics*), joiden avulla tieto-

kannan rakenne ja merkitys voidaan parhaiten säilyttää, jotta tietokannasta saadaan talteen eheä, autenttinen kopio. Wilson (2008, 15) määrittelee keskeiset ominaisuudet seuraavasti:

> ”The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record." (Wilson 2008, 15.)

Digitaalisten asiakirjojen keskeisten ominaisuuksien valinnasta on kiistelty tiedeyhteisön sisällä paljon viimeisen vuosikymmenen aikana. Fanielin ja Yakelin vuonna 2011 tekemän kirjallisuuskatsauksen mukaan keskeisten ominaisuuksien valintaperusteet voidaan jakaa karkeasti kahteen luokkaan: 1) keskeiset tekniset ominaisuudet aineiston esittämiseksi alkuperäistä vastaavassa muodossa ja 2) keskeiset ominaisuudet aineiston tietosisällön merkityksen säilyttämiseksi. Keskeisiksi ominaisuuksiksi on alan kirjoituksissa katsottu muun muassa digitaalisen objektin toiminnallisuus, ulkoasu, käyttökokemus, käyttöoikeudet, käyttötarkoitus, ja monia muita ominaisuuksia. Koska keskeisiksi katsottujen ominaisuuksien kirjo on näin laaja, ei kaikkia näitä ominaisuuksia voida esittää metatiedon avulla. Digitaaliseen aineistoon liittyy paljon ominaisuuksia, ja se, kuinka keskeisenä kutakin ominaisuutta pidetään, riippuu käyttäjästä, sekä siitä käyttötarkoituksesta, jota varten aineistoa säilytetään. (Faniel & Yakel 2011, 2–3, 7–8.)

Säilytyksen kannalta keskeiset relaatiotietokantojen ominaisuudet pyrittiin määrittelemään Euroopan unionin rahoittaman *E-ARK (European Archival Records and Knowledge Preservation)* -projektin yhteydessä (Ferreira 2016, 18). Lukuisien yksityisten ja julkisten instituutioiden, sekä eri säilytysformaattien parissa työskentelevien tahojen avustuksella päädyttiin koostamaan seuraavanlainen keskeisten ominaisuuksien lista:

- *Tietosisältö*. Tietokannan sisältämä data, eli tietokannan solujen arvot, tulee säilyttää sellaisenaan. Tietotyyppien mahdollinen vaihtuminen konversiossa ei saa johtaa tiedon menettämiseen.
- *Relaationaalinen rakenne*. Kaikki se tieto, joka liittyy tietokannan relaationaalisen rakenteen säilymiseen, tulee säilyttää. Tähän kuuluvat tietokannan skeemat, taulut, avaimet, sekä attribuutit ja niiden tietotyypit.

- *Käyttäytymiseen liittyvät tiedot*, jotka auttavat ymmärtämään sitä, kuinka tietokanta on toiminut. Näihin kuuluvat tietokannan käyttäjät, roolit ja käyttöoikeudet, näkymät, herättimet, tallennetut proseduurit ja funktiot, sekä tietokannan rajoitteet.
- *Kuvailutiedot*, jotka kuvaavat tietokannan eri ominaisuuksia, kuten sen kenttiä ja tietotyyppejä.
- *Tietokantasovelluksen dokumentaatio*, josta käy ilmi, millaisessa vuorovaikutuksessa tietokantasovellus ja tietokanta ovat olleet.

  (Ferreira 2016, 18–19.)

Nämä keskeiset ominaisuudet antavat viitteitä siitä, mitkä tietokannan ominaisuudet on ainakin hyvä säilyttää, jotta tietokannan säilyminen eheänä ja autenttisena voidaan varmistaa. Freitasin (2012, 31) mukaan keskeiset ominaisuudet tulee valita siten, että tietokannan alkuperäinen merkitys säilyy sellaisena kuin se oli alun perin tarkoitettu. Tietokannan tulee säilyttää ”hyväksyttävä alkuperäisyyden taso”. Säilytetyn jäljennöksen täytyy kyetä toimimaan evidenssinä siitä tietokannasta, joka se on joskus ollut. Tähän tavoitteeseen voidaan päästä vain analysoimalla jokainen arkistoitava kohde tapauskohtaisesti ja perinpohjaisesti, jotta voidaan päättää, minkä ominaisuuksien säilyttäminen palvelee tarkoitusta parhaiten. (Freitas 2012, 31.)

2.5 Tietokantojen pitkäaikaissäilytyksen vaihtoehdot

Tietokannat eivät ole yksiselitteisiä, määrämuotoisia dokumentteja, joiden sisältö voidaan yksinkertaisesti kopioida kiintolevyltä toiselle. Ne ovat interaktiivisia, dynaamisia järjestelmiä, jotka jatkuvasti kasvavat ja päivittyvät. Tietokannan sisältö ei ole käsitteellistettävässä muodossa ilman asianmukaisen tietokannanhallintajärjestelmän tai muun tietokannan käyttämiseksi rakennetun sovelluksen apua. Relaatiotietokannoissa tieto on pilkottu pieniin osiin ja hajautettu tietokannan eri taulujen sisälle. Se kootaan yhteen vasta tietokantasovelluksen käyttäjälle tai ylläpitäjälle tuottamaa näkymää varten, eikä kaikkea tietokannan sisältämää tietoa välttämättä koskaan tulosteta näytölle lainkaan. Lisäksi eri tietokantajärjestelmien koko voi vaihdella hyvinkin paljon tietokannan tyypin, tietosisällön ja käyttötarkoituksen mukaan. Kaikki nämä ominaisuudet tekevät relaatiotietokantojen säilyttämisestä monimutkaisemman operaation kuin muiden digitaalisten asiakirjojen, kuten teksti- tai kuvatiedostojen, säilyttäminen.

Säilytysstrategian valintaan vaikuttaa säilytettävän tietokannan luonne. Shepherd ja Smith (2000, 58–60) ovat jaotelleet tietokannat niiden käyttötavan perusteella avoimiin ja suljettuihin, sekä dynaamisiin ja staattisiin tietokantoihin. Avoimilla tietokannoilla tarkoitetaan sellaisia tietokantoja, joihin tietoa voidaan vielä lisätä, ja vastaavasti suljetuilla sellaisia tietokantoja, joihin ei lisäyksiä tai muutoksia enää tule. Staattisilla tietokannoilla puolestaan tarkoitetaan tietokantaa, johon kerran syötettyä tietuetta ei enää jälkeenpäin muuteta, ja dynaamisilla tietokannoilla tietokantaa, jonka tietueita voidaan jatkuvasti päivittää tai poistaa. (Shepherd & Smith 2000, 58–60.)

Säilytyksen kannalta suljetut, staattiset tietokannat ovat yksinkertaisimpia. Ne voidaan säilyttää sellaisenaan, sillä tiedetään, että mitään tietokannan sisältämää tietoa ei ole muutettu, eikä uutta tietoa ole enää tietokantaan tulossa. Tällaisia tietokantoja muodostavat esimerkiksi yksittäiset, jo päättyneet lomakekyselyt. Myös avoimet, staattiset tietokannat ovat melko suoraviivaisia säilytettäviä. Tietokannasta voidaan ottaa aika ajoin *tilannekuva* eli *otos* (engl. *snapshot*), joka siirretään säilytysmedialle. Koska tietueiden sisältöä ei muuteta eikä tietueita poisteta missään tietokannan elinkaaren vaiheessa, saadaan arkistointiprosessissa talteen kaikki tietokannan sisältämä tieto. Suljetut, dynaamiset tietokannat ovat yleensä jonkin organisaation jälkeensä jättämiä tietokantoja, joita ei enää päivitetä, jolloin ne voidaan säilyttää sellaisenaan. (Ashley 2004, 67–68.)

Avoimet, dynaamiset tietokannat ovat säilytyksen kannalta ongelmallisia. Tällaisia tietokantoja ovat esimerkiksi pankin tilitiedot, jotka jatkuvasti päivittyvät. Tietokannan otoksen tallentaminen säännöllisin väliajoin ei tarjoa kattavaa kuvaa tietokannan tapahtumista tietyllä aikavälillä, sillä tietokannan sisältämä tieto saattaa muuttua paljonkin ajanhetkestä toiseen. Juhlapyhien ja joulun alla tapahtuu enemmän transaktioita kuin muihin vuodenaikoihin, ja ihmisillä on vähemmän rahaa tileillään. Vastaava trendi on havaittavissa kuukauden alku- ja loppupuolella. Suljetussa dynaamisessa tietokannassa oleva sisältö ei tämän vuoksi välttämättä anna hyvää kuvaa tyypillisestä tilanteesta, sillä lopullinen tilanne kuvaa vain tietokannan viimeisen käyttöajankohdan tilanteen. (Ashley 2004, 67–68.)

Vuoteen 2000 mennessä oli tunnistettu kolme erilaista strategiaa tietokantojen pitkäaikaissäilytykseen. Nämä ovat alkuperäisen *teknologian säilyttäminen (*engl. *technology preservation)*; alkuperäisen teknologian *emuloiminen* (engl. *emulation*), sekä tietokannan konvertoiminen toiseen muotoon. (ks. mm. van Horik & Roorda 2009.) Nämä säily-

tysstrategiat eroavat toisistaan tekniseltä toteutukselta hyvin paljon, ja johtavat säilytyksen kannalta erilaisiin lopputuloksiin. Jos tietokannan toiminnallisuus halutaan säilyttää, tulee nykyteknologian puitteissa kyseeseen joko teknologian säilyttäminen, jolloin alkuperäinen laitteisto, ohjelmisto ja tietokantajärjestelmä säilytetään kokonaisuudessaan, tai tietokantajärjestelmän emuloiminen toisen järjestelmän sisällä, jolloin ainoastaan tietokantajärjestelmä säilytetään. Nämä lähestymistavat soveltuvat käyttöön silloin, kun on tärkeää säilyttää alkuperäinen käyttöympäristö ja käyttökokemus mahdollisimman lähellä alkuperäistä.

Teknologian säilyttämisellä tarkoitetaan alkuperäisen laitteiston, sekä käytössä olleen käyttöjärjestelmän, tietokannanhallintajärjestelmän ja tietokantasovelluksen säilyttämistä sellaisenaan. Tällöin tulee säilyttää myös asiaankuuluvat ohjekirjat, sekä tietokantajärjestelmän dokumentaatio, jotta järjestelmää pystytään käyttämään myöhemminkin. Teknologian säilyttäminen on kuitenkin monimutkainen ja kallis operaatio, joka vaatii paljon tilaa, minkä vuoksi sitä ei pidetä realistisena ratkaisuna digitaalisen tiedon pitkäaikaissäilytykselle. Järjestelmää ei voida myöskään ylläpitää loputtomiin, sillä ajan kuluessa laitteisto ja tallennusmedia rappeutuvat ja lakkaavat toimimasta, jollei varaosia enää valmisteta. (Lee ym. 2002, 95.)

Emulaation avulla voidaan vanhaksi jäänyttä laitteisto- tai ohjelmistoympäristöä jäljitellä siten, että vanhentunut ohjelmisto voidaan suorittaa uudessa laite- tai ohjelmaympäristössä. Emulaation vahvuutena pitkäaikaissäilytyksen kannalta on se, että alkuperäisessä tiedostomuodossa olevia asiakirjoja voidaan lukea ilman, että niitä joudutaan konvertoimaan toisiin muotoihin. Vanhentunut ohjelma voidaan suorittaa sellaisenaan, jolloin myös tiedon alkuperäinen esitysmuoto säilyy. Emulaatiota suositellaan tietokantojen säilytysstrategiaksi silloin, kun itse tietokantasovellus, tai jokin sen toiminnallisuudessa, koetaan säilyttämisen kannalta merkittäväksi; silloin, kun on tärkeää säilyttää alkuperäinen käyttökokemus sellaisenaan; sekä silloin, kun kyseessä on puutteellisissa olosuhteissa arkistoitu tietokanta, jonka käyttämiseen tarvitaan tietty ohjelmisto. (van Essen ym. 2011, 14.) Käytännön sovelluksia tietokantojen pitkäaikaissäilytyksestä emulaation avulla on olemassa vielä vähän, ja emulaatiota pidetään toistaiseksi liian kalliina ja monimutkaisena tekniikkana useimpien muisti-instituutioiden tarpeisiin (ks. Delve ym. 2014).

Tietokannan konversiossa tietokanta säilytetään erillään alkuperäisestä tietokannanhallintajärjestelmästä, ja pyritään muuntamaan mahdollisimman neutraaliin, pitkäikäiseen

tiedostomuotoon. Alkuperäistä tietokantajärjestelmää ja tietokannan toiminnallisuutta ei pyritä säilyttämään, vaan tietokannan rakenne ja tietosisältö pyritään muuntamaan sellaiseen säilytysformaattiin, jossa se on mahdollisimman lähellä alkuperäistä muotoa. Konversion tarkoituksena on muuntaa tietokanta vanhentumassa olevasta tiedostomuodosta ajankohtaisempaan muotoon, ja siten varmistaa pääsy tietokannan tietosisältöön pitkälle tulevaisuuteen, vaarantamatta kuitenkaan tietokannan eheyttä tai autenttisuutta.

Perinteisesti relaatiotietokantojen arkistoinnissa on turvauduttu konversiomenetelmiin, joissa tietokanta muunnetaan johonkin yksinkertaisempaan muotoon, kuten pilkuilla erotetuksi peräkkäistiedostoksi. Tällöin tietokannan normalisointi joudutaan kuitenkin purkamaan, jotta tietokanta saadaan arkistoitavaan muotoon, mikä vähentää yksittäisten taulujen määrää. Normalisoinnin purkaminen aiheuttaa tietokannan alkuperäisen rakenteen ja tietueiden välisten relaatioiden rikkoutumisen. Tässä muodossa tallennettu tietokanta ei ole myöskään enää palautettavissa tietokannanhallintajärjestelmään, jolloin tiedon käsittelymahdollisuudet menetetään. Tietokantojen konversiota yksinkertaisempaan tiedostomuotoon ei nykyisin suositella lainkaan, sillä siihen sisältyy suuri riski siitä, että tietoa kadotetaan prosessissa. (Digital Preservation Testbed 2003, 22.)

Tiettyyn ohjelmistovalmistajaan sidottuja tiedostomuotoja ei myöskään nykyisin suositella säilytysmuodoksi, sillä niillä on suuri riski vanhentua ja poistua käytöstä muutamassa vuosikymmenessä (ks. mm. Lawrence 2001). Tietokannan säilyttäminen tietokannanhallintajärjestelmän omassa tallennusmuodossa ja sen konvertoiminen aika ajoin uuteen ohjelmistoversioon voi kuitenkin olla hyödyllinen ratkaisu silloin, kun säilytysaika on suhteellisen lyhyt – enintään 5-10 vuotta. Tietokannan päivittäminen uuteen versioon ei yleensä sisällä suuria riskejä, eikä vaaranna tietokannan eheyttä ja autenttisuutta. Konversio uuteen versioon joudutaan kuitenkin tekemään uudelleen muutaman vuoden välein, aina uuden version ilmestyessä. Mitä enemmän konversioita joudutaan tekemään – ja mitä kauemmas alkuperäisestä versiosta tullaan – sitä suurempi on riski siitä, että tietoa kadotetaan prosessissa. (Digital Preservation Testbed 2003, 21.)

2.6 XML tietokantojen säilytyksessä

Tänä päivänä vakiintunut käytäntö tietokantojen säilyttämiseksi pitkällä aikavälillä on tietokannan konvertoiminen johonkin avoimeen, standardoituun tiedostomuotoon. Avoimen, standardoidun säilytysmuodon ajatellaan vähentävän konversioiden tarvetta

tulevaisuudessa. Voidaan myös olettaa, että laajassa käytössä olevat, avoimet standardit pysyvät käytössä suljettuja, tiettyyn ohjelmistovalmistajaan sidottuja tiedostomuotoja pidempään. On todennäköistä, että myös avoimet, standardoidut tiedostomuodot tulevat muuttumaan, mutta muutoksen arvellaan olevan hitaampi kuin kaupallisilla tiedostomuodoilla. (ks. mm. van Horik & Roorda 2009, 195; Rahman ym. 2015, 255.)

XML on tällainen avoin ja neutraali tiedostomuoto, jota on käytetty jo 1990-luvun lopulta lähtien tiedon kuvaamiseen ja tiedonvaihtoon. XML on avoimen standardin tekstipohjainen kuvauskieli, joka kehitettiin erityisesti rakenteisten dokumenttien kuvailun tarpeisiin standardoidun kuvauskielen, *SGML*:n *(Standard Generalized Markup Language)*, pohjalta (W3C 2008). XML:ssä on pyritty täsmälliseen, yleiskäyttöiseen formaattiin, jota on helppo tulkita ja käyttää, ja jonka eheys on helppo varmistaa ohjelmallisesti. XML on digitaalisessa pitkäaikaissäilytyksessä yleisesti käytetty tiedostomuoto muun muassa metatiedon esittämiseen, säilytettävän aineiston *kapseloimiseen* (engl. *encapsulation*), sekä itsenäisenä *säilytysformaattina* (engl. *preservation file format*), johon aineistoa voidaan konvertoida (Becker ym. 2008, 2939).

XML-dokumentti koostuu *elementeistä* (engl. *element*) ja niitä tarkentavista *attribuuteista* (engl. *attribute*), joita kuvataan niin kutsutuilla *tägeillä* (engl. *tag*) (ks. kuvio 5). XML-dokumentin elementit muodostavat puurakenteen, ja jokaisella XML-dokumentilla on oltava yksi juurielementti. XML-dokumentin rakenne ja sallitut elementit kuvaillaan erillisessä *DTD* (*Document Type Definition*) -dokumentissa tai *XML-skeemassa* (engl. *XML schema*). XML-dokumenttien käsittelyyn on tarjolla paljon erilaisia työkaluja, kuten *XPath*, jonka avulla XML-dokumentista voidaan osoittaa tiettyjä kohtia, sekä *XQuery*-kyselykieli, joka on suppea, SQL-kieltä muistuttava kyselykieli. (W3C 2008; W3C 2016; W3C 2017.)

```
<elementti attribuutti=”arvo”>
</elementti>
```

Kuvio 5: XML-muotoinen tägi.

XML:n eduiksi pitkäaikaissäilytyksessä katsotaan se, että XML:n avulla voidaan asiakirjan konteksti, sisältö ja rakenne dokumentoida yhdessä ja samassa tiedostossa. XML ei myöskään ole sidoksissa tiettyyn laitteisto- tai ohjelmistoympäristöön, mikä tekee siitä monia kaupallisia tiedostomuotoja houkuttelevamman ratkaisun digitaalisen säilytyk-

sen tarpeisiin. Tiedon muuntaminen valmistajasta riippumattomaan, avoimeen muotoon helpottaa tiedonsiirtoa eri järjestelmien välillä, ja vähentää säilytysformaatin vanhenemisen riskiä. XML:n eduiksi voidaan nähdä myös se, että sitä on helppo käsitellä ohjelmallisesti. (Digital Preservation Testbed 2003, 24.)

Tässä tutkielmassa keskitytään tarkastelemaan XML:n käyttöä relaatiotietokantojen pitkäaikaissäilytyksessä. Tietokantojen pitkäaikaissäilytyksessä XML:ää käytetään tyypillisesti säilytysformaattina, johon tietokannan rakenne, sisältö ja mahdolliset muut ominaisuudet konvertoidaan. Tietokannan rakenne voidaan kuvailla eksplisiittisesti XML-skeeman tai DTD-dokumentin avulla, minkä ansiosta XML soveltuu erinomaisesti tietokantojen säilytykseen (Digital Preservation Testbed 2003, 24). XML-konversio soveltuu parhaiten suljettujen, staattisten tietokantojen säilyttämiseen sekä tietokantaotosten säilyttämiseen. Useimmat XML-konversioon pohjautuvat säilytysratkaisut tukevat myös arkistoidun tietokannan palauttamista toiminnassa olevaan tietokannanhallintajärjestelmään, mikä mahdollistaa kyselyjen tekemisen säilytettyyn tietokantaan. (Freitas 2012, 36.) Seuraavassa luvussa tutustaan tarkemmin XML-konversioon pohjautuvien säilytysmenetelmien kehitykseen ja alan tutkimuksen nykytilanteeseen.

3 XML-KONVERSIO SÄILYTYSRATKAISUNA

XML:n mahdollisuuksia relaatiotietokantojen pitkäaikaissäilytyksessä on tutkittu 2000-luvun vaihteesta lähtien lukuisien eri tahojen toimesta eri puolilla maailmaa. Tutkimukselle pontta on antanut alati kasvava tiedon määrä sekä tietokantojen yhä suurempi merkitys organisaatioiden jokapäiväisessä toiminnassa. Tietokantojen pitkäaikaissäilytykseen liittyvä tutkimus on ollut kuitenkin suhteellisen vähän esillä informaatiotieteen ja tietojenkäsittelytieteen julkaisuissa siihen nähden, kuinka suuri merkitys tietokannoilla on yritysten ja organisaatioiden toiminnassa tänä päivänä. Aihealueen tutkimusta on esitelty lähinnä kansainvälisissä työpajoissa ja konferensseissa. Myös muutamia väitöskirjoja on maailmalla aiheesta kirjoitettu. Kotimaista tutkimusta tietokantojen pitkäaikaissäilytykseen liittyen on julkaistu hyvin vähän, vaikka pitkäaikaissäilytysratkaisuja on tutkittu ja kehitetty niin Kansallisarkiston toimesta kuin monien yksityisten tutkimuslaitosten ja organisaatioidenkin toimesta.

Kansainvälinen tutkimus on keskittynyt XML-pohjaisten säilytysformaattien tekniseen toteutukseen sekä konversiotyökalujen kehittämiseen. Tavoitteena on ollut tuottaa työkaluja, joiden avulla viranomaiset voivat itse koostaa tietokannasta säilytettävän kokonaisuuden. XML-konversioon pohjautuvat säilytysmenetelmät eivät pyri säilyttämään tietokannan toiminnallisuutta, käyttöympäristöä tai alkuperäistä tietokantasovellusta sellaisenaan. Painotus on ollut tietokannan keskeisten, eheyden ja autenttisuuden takaavien ominaisuuksien säilyttämisessä. Tässä luvussa tarkastellaan XML-konversioon perustuvien relaatiotietokantojen säilytysmenetelmien tutkimusta ja kehityssuuntauksia 2000-luvulta lähtien.

3.1 XML-konversioon perustuvien säilytysmenetelmien kehitys

Alankomaiden kansallisarkisto oli ensimmäisiä XML:n käyttöä relaatiotietokantojen pitkäaikaissäilytyksessä tutkineita tahoja. Vuonna 2000 alkaneessa Digital Preservation Testbed -projektissa kartoitettiin erilaisia vaihtoehtoja digitaalisten aineistojen säilyttämiseksi. Projektin tarkoituksena oli tuottaa pitkäaikaissäilytysratkaisuja Alankomaiden julkishallinnon tarpeisiin. Projektin lähtökohtana oli määritellä, mitkä pitkäaikaissäilytysratkaisut tai niiden yhdistelmät soveltuisivat tarkoitukseen parhaiten. Projektin aikana toteutettiin testialusta, jonka avulla eri säilytysmenetelmien soveltuvuutta julkishal-

linnon tarpeisiin testattiin. Tutkimus kohdistui erityisesti kolmeen menetelmään: migraatioon, XML-konversioon, sekä emulointiin. Projektin aikana arvioitiin muun muassa eri menetelmien tehokkuutta ja kustannuksia, sekä niiden mahdollisuuksia ja rajoituksia. Tietokantojen pitkäaikaissäilytystä tarkasteltiin omana osa-alueenaan. (Digital Preservation Testbed 2003, 4.)

Digital Preservation Testbed -projektissa XML todettiin erinomaiseksi vaihtoehdoksi tietokantojen pitkäaikaissäilytykseen, ja projektissa päädyttiin suosittelemaan tätä ratkaisua Alankomaiden kansallisarkiston tarpeisiin. (Digital Preservation Testbed 2003, 26) Digital Preservation Testbed tuotti lisäksi suosituksen siitä, mistä osista tietokannan säilytystiedoston tulee koostua. Työryhmän mukaan tietokannan XML-muotoon konvertoidun rakenteen ja tietosisällön ja rakenteen säilytystiedostossa tulee säilyttää metatietoa tietokannan käyttökontekstista, tietokantasovelluksen dokumentaatio SQL-kyselyineen ja kuvakaappauksineen, alkuperäinen tietokantatiedosto sellaisenaan, sekä lokitiedosto säilytyksen aikaisista toimista, jotta tietokannan autenttisuudesta voidaan varmistua. (Digital Preservation Testbed 2003, 36–39.)

3.1.1 MIXED

Alankomaiden tieteellisen data-arkiston *(Data Archiving and Networked Services)* vuonna 2007 aloittama MIXED *(The Migration to Intermediate XML for Electronic Data)* -projekti jatkoi siitä, mihin Digital Preservation Testbed -projektissa jäätiin. MIXED-projektin tarkoituksena oli kehittää käytäntöjä ja työkaluja digitaalisen tiedon pitkäaikaissäilytyksen tarpeisiin Alankomaiden tieteellisessä data-arkistossa. Projektin aikana kehitettiin SDFP *(Standard Data Formats for Preservation)* -niminen XML-skeema laskentataulukoiden, tilastollisen datan ja relaatiotietokantojen säilyttämiseen. SDFP-dokumentti toimi kokoavana säilytystiedostona, johon tallennettiin metatietoa tietokannan alkuperästä ja konversioprosessista, sekä säilytettävän taulukon tai tietokannan sisältö omana XML-dokumenttinaan. Tarkoituksena oli, että SDFP toimisi vain säilytysformaattina, ja tarvittava asiakirja palautettaisiin haluttuun tiedostoformaattiin jakelua varten. (van Horik & Roorda 2011.)

Van Horik ja Roorda eivät kuvaa artikkelissaan tietokantojen XML-pohjaista säilytysmuotoa yksityiskohtaisesti (ks. van Horik & Roorda 2011). MIXED-projektin dokumentaatiosta vuodelta 2010 käy kuitenkin ilmi, että ohjelmiston käyttämä tietokantojen

säilytysmuoto perustui alun perin Antverpenin arkiston kehittämään *eDAVID*-tiedostomuotoon (Data Archiving and Networked Services (DANS) 2010). MIXED-projekti päättyi vuonna 2010, eikä ohjelmisto ole enää käytettävissä. Alankomaiden tieteellinen arkisto on sittemmin ottanut SIARD-säilytysformaatin käyttöön relaatiotietokantojen pitkäaikaissäilytysmuotona (Data Archiving and Networked Services (DANS) 2015).

3.1.2 SIARD

Samoihin aikoihin Digital Preservation Testbedin kanssa oli Sveitsissä käynnissä ARELDA *(Archiving of Electronic Data and Records)* -projekti, jonka tarkoituksena oli tuottaa pitkäaikaissäilytysratkaisuja Sveitsin kansallisarkiston tarpeisiin. Eräs projektin tavoitteista oli tuottaa tehokas, automaattinen ja standardoitu ratkaisu relaatiotietokantojen arkistointiin.(Swiss Federal Archives 2001, 1, 5.) Projektin aikana kehitettiin SIARD *(Software Independent Archiving of Relational Databases)* -säilytysformaatti, joka on tällä hetkellä laajimmalle levinneitä ja tunnetuimpia ratkaisuja relaatiotietokantojen pitkäaikaissäilytykseen. SIARD on avoimen lähdekoodin säilytysformaatti, joka perustuu avoimiin standardeihin, kuten Unicode, XML ja SQL:1999, sekä ZIP64-tiedostoformaattiin, johon kaikki tieto tallennetaan pakkaamattomana (Heuscher ym. 2004, 7).

SIARD:n ensimmäinen versio sai Sveitsissä kansallisen standardin vuonna 2013. Nykyisin käytössä on kehittyneempi versio, SIARD 2.0, joka kehitettiin Euroopan komission rahoittaman E-ARK-projektin yhteydessä. Se sai Sveitsissä kansallisen standardin vuonna 2016 (eCH-0165 2016). Myös Tanskan kansallisarkisto on kehittänyt SIARD:sta oman versionsa, joka kantaa nimeä SIARDDK. SIARDDK kehitettiin nimenomaan Tanskan kansallisarkiston tarpeisiin, ja eroaa SIARD 1.0:sta muun muassa kansiohierarkioiden ja tiedostojen sijainnin suhteen. SIARD on levinnyt laajaan käyttöön eri puolille Eurooppaa. Sveitsin ja Tanskan lisäksi se on käytössä muun muassa Hollannin, Saksan ja Ranskan kansallisarkistoissa tietokantojen säilytysformaattina (Das Bundesarchiv 2011; Data Archiving and Networked Services (DANS) 2015; CINES 2017).

Suomessa SIARD:in käyttöönottoa on tutkittu muun muassa Opetus- ja kulttuuriministeriön Avoin tiede ja tutkimus -hankkeen yhteydessä. Hankkeen aikana tutkittiin muun

muassa tutkimusaineistojen tiedostomuotoja ja niiden pitkäaikaissäilytyskelpoisuutta. Selvityksen loppuraportissa todettiin SIARD-muoto lupaavaksi tekniikaksi relaatiotietokantojen säilytykseen. (Avoin tiede ja tutkimus 2017, 34.)

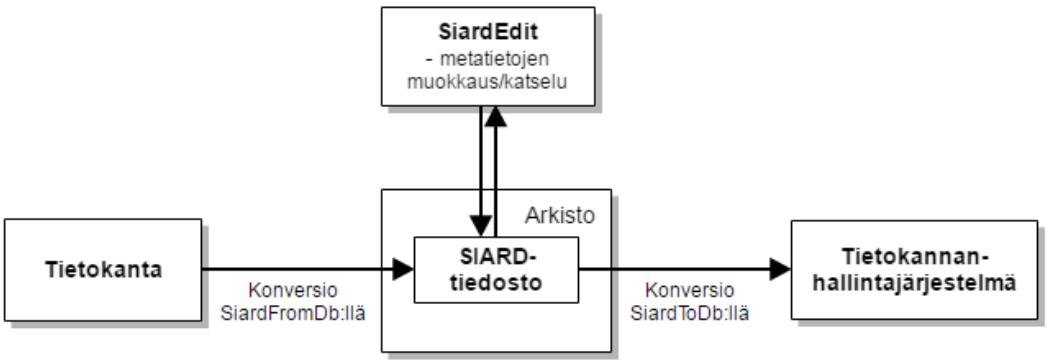
SIARD tallentaa kaiken tietokantaan liittyvän tiedon XML-tiedostoina, jotka on koottu yhteen ZIP64-muotoiseen pakettiin. SIARD-tiedosto on jaettu kahteen kansioon, *header* ja *content,* kuvion 6 mukaisella tavalla. Header-kansiossa sijaitseva *metadata.xml*-dokumentti sisältää arkistoidun tietokannan rakenteen ja kuvauksen. Tietokannan sisältö puolestaan tallennetaan content-kansioon siten, että kullekin tietokannan taululle luodaan oma XML-dokumenttinsa. Content-kansion sisältö on kuitenkin riippuvainen header-kansion sisällöstä, minkä vuoksi arkistoidun tietokannan sisältöä ei voi selailla tavallisella XML-lukijaohjelmalla, vaan sisällön tarkastelemiseksi tulee käyttää erityistä, tarkoitusta varten kehitettyä ohjelmaa. (Rahman ym. 2012, 496-497.)

```
/ ..................................................... (ZIP-paketin juuri)
└── header/ ..................................... (Tietokannan metatiedot)
    └── metadata.xml
    └── metadata.xsd
    └── version/
        └── 2.0/ ................................ (SIARD:n versionumero)
└── content/ ..................................... (Tietokannan sisältö)
    └── schemaM/ ........ (Tietokannan skeema, jossa M on yksilöivä tunniste)
        └── tableN/ ....... (Tietokannan taulu, jossa N on yksilöivä tunniste)
            └── tableN.xml
            └── tableN.xsd
```

Kuvio 6: SIARD 2.0 -hakemistorakenne.

Myös tietokannan kuvaileva, rakenteellinen ja tekninen metatieto tallennetaan *metadata.xml* -tiedostoon. Metatiedon avulla kuvaillaan ja dokumentoidaan muun muassa tietokannan alkuperäinen käyttötarkoitus ja tekninen ympäristö, tietokannan arkistointiprosessi ja siitä vastannut henkilö, sekä kukin tietokantaan sisältyvä taulu erikseen. (Ferreira 2016, 33.) SIARD-pakettiin tallennetaan lisäksi tietokantaan liittyvä dokumentaatio, josta käy ilmi muun muassa käytetty tietokantajärjestelmä, tietokannan elementtien väliset suhteet, sekä kuvaukset tietokannan sisältämästä datasta. Dokumentaatioon liitetään tietokantasovelluksen käyttöohjeet kuvakaappauksineen. (Heuscher ym. 2004, 7-15.)

SIARD Suite on Sveitsin kansallisarkiston kehittämä ohjelmisto, jonka avulla relaatiotietokannan konversio SIARD-säilytysmuotoon suoritetaan. SIARD Suite mahdollistaa erityyppisten metatietojen syöttämisen, tietokannan muuntamisen eri tietokantajärjestelmien välillä, sekä tietokannan palauttamisen toiminnalliseen järjestelmään, jolloin siihen voidaan tehdä hakuja. (Heuscher ym. 2004, 7-15; Swiss Federal Archives 2010.)



Kuvio 7: Tietokannan arkistointi SIARD Suiten avulla.

SIARD Suite sisältää kolme eri sovellusta tietokannan arkistointiprosessin eri vaiheisiin. Ensimmäisessä vaiheessa arkistoitava tietokanta analysoidaan ja konvertoidaan tietokannanhallintajärjestelmästä SIARD-muotoon *SiardFromDb*-nimisen työkalun avulla. Työkalu luo tietokannan pohjalta XML-tiedostot, jotka kootaan pakkaamattomaan ZIP64-kansioon. Toisessa vaiheessa XML-tiedostot siirretään *SiardEdit*-kuvailusovellukselle, jonka avulla käyttäjä liittää SIARD-tiedostoon teknistä, kuvailevaa ja kontekstista kertovaa metatietoa tietokannan eri hierarkiatasoilla. SiardEditin avulla voidaan myös tarkastella arkistoidun tietokannan sisältöä. Tässä vaiheessa pakettiin liitetään järjestelmädokumentaatio, joka voi olla PDF- tai TIFF-muodossa. Dokumentaatio voi sisältää muun muassa käyttöohjeita, lokitiedostoja ja tietoturvaraportteja. Sovellus liittää metatiedot pakettiin, joka on nyt valmis arkistoitavaksi. SIARD:in kolmas työkalu, *SiardToDb*, on tarkoitettu arkistoidun tietokannan palauttamiseksi haluttuun tietokannanhallintajärjestelmään. Ilman tätä vaihetta ei arkistoituun tietokantaan voida tehdä kyselyjä. SIARD Suiten käyttövaiheet on havainnollistettu kuviossa 7. (Heuscher ym. 2004, 7-15; Swiss Federal Archives 2010.)

3.1.3 RODA

Myös Portugalin kansallisarkistolla *(Arquivo Nacional da Torre do Tombo)* oli kehitteillä oma pitkäaikaissäilytysratkaisunsa 2000-luvun loppupuolella. Vuonna 2006 alkaneessa yhteisprojektissa Minhon yliopiston kanssa toteutettiin RODA *(Repository of Authentic Digital Objects)* -niminen digitaalinen pitkäaikaissäilytysjärjestelmä. RODA:n tavoitteena oli tuottaa kokonaisvaltaiset digitaalisen säilytyksen ratkaisut Portugalin julkishallinnolle. RODA perustuu avoimen lähdekoodin standardeihin ja on yhteensopiva OAIS-viitemallin kanssa. RODA:a ei ole tarkoitettu vain relaatiotietokantojen säilyttämiseen, vaan se tukee myös tekstidokumenttien, kuvien, videoiden ja äänitiedostojen säilytystä. Nykyisin RODA:a kehittää KEEP SOLUTIONS -niminen yritys. (Ramalho ym. 2008; RODA 2016.)

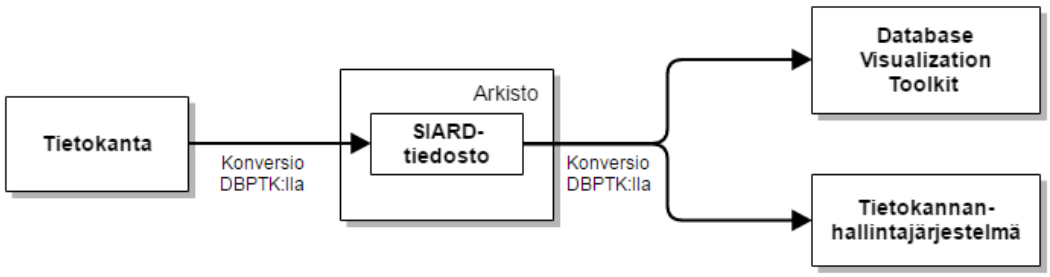
Alun perin relaatiotietokantojen säilytys RODA:ssa toteutettiin *DBML (Database Markup Language)* -formaatin avulla. DBML on XML-pohjainen kuvauskieli, joka oli suunniteltu relaatiotietokantojen konvertoimiseen eri tietokantajärjestelmien välillä. RODA-projektin yhteydessä tekniikka otettiin käyttöön relaatiotietokantojen pitkäaikaissäilytysformaattina. DBML on SIARD-säilytysformaattia yksinkertaisempi tiedostoformaatti. Tietokannan rakenne ja data sisällytetään yhteen ja samaan DBML-dokumenttiin, joka voidaan edelleen paketoida OAIS-viitemallin mukaiseen siirtopakettiin. Nykyisin RODA-pitkäaikaissäilytysjärjestelmä tukee myös SIARD-säilytysformaattia. (Ramalho ym. 2008; Ramalho ym. 2014.)

3.1.4 Database Preservation Toolkit

Database Preservation Toolkit (jatkossa: DBPTK) on RODA-projektista irralliseksi projektikseen lähtenyt ohjelmisto, jonka kehitystä jatkettiin Euroopan Komission rahoittaman E-ARK-projektin yhteydessä. DBPTK on työkalu, joka mahdollistaa relaatiotietokantojen konversion SIARD- ja DBML-muotoihin. Ohjelmisto mahdollistaa myös tietokannan konversion tietokantajärjestelmästä toiseen, sekä arkistoidun tietokannan palauttamisen toiminnassa olevaan tietokannanhallintajärjestelmään, jolloin siihen voidaan tehdä kyselyjä. (Ramalho ym. 2014; Ferreira ym. 2016.)

DBPTK tukee tietokantojen muuntamista SIARD 1.0, SIARD 2.0 sekä SIARDDK-muotoihin. Sovelluksessa oli alun perin tuki vain tietokannan konvertoimiseksi DBML-muotoon. E-ARK-projektin yhteydessä ohjelmistoon rakennettiin tuki SIARD-säilytysformaatille, sillä se sopi paremmin projektin tarpeisiin. SIARD:n eduiksi nähtiin se, että tallentaa tietokannan rakenteen ja tietosisällön eri tiedostoihin, sekä se, että SIARD tukee näkymien ja tallennettujen proseduurien säilyttämistä, toisin kuin DBML.(Ramalho ym. 2014.)

DBPTK:n avulla arkistoitu SIARD-tiedosto voidaan konvertoida toimivaan tietokannanhallintajärjestelmään kyselyjen tekemiseksi, tai avata erilliseen, *Database Visualization Toolkit* -nimiseen katseluohjelmaan tarkasteltavaksi. (Ferreira 2016, 25.) Tietokannan arkistointiprosessi DBPTK:n avulla on visualisoitu kuviossa 8.



Kuvio 8: Tietokannan arkistointi DBPTK:n avulla.

3.2 Tietovarastoteknologia relaatiotietokantojen säilytyksessä

Rahman ym. (2015) ovat tutkineet tietovarastoissa käytettävän teknologian soveltamista relaatiotietokantojen pitkäaikaissäilytykseen INESC Porton, Minhon yliopiston, Portugalin kansallisarkiston sekä tiede- ja teknologiayhdistys FCT:n rahoittamassa DBPreserve-projektissa. Säilytettävä tietokanta muunnettiin relaatiomallista moniulotteiseen tietomalliin purkamalla tietokannan normalisointi, mikä yksinkertaistaa tietokannan rakennetta ja parantaa kyselyjen tehokkuutta. (Rahman ym. 2015, 116–117.) Säilytysformaatin pohjana käytettiin SIARD-säilytysformaattia, jota laajennettiin liittämällä SIARD-tiedostoon moniulotteisen tietomallin kuvaavaa metatietoa (ks. kuvio 9). (Rahman ym. 2015, 121.)

```
/ ..... (ZIP-paketin juuri)
└── header/ ..... (Tietokannan metatiedot)
    └── dw.xml
    └── dw.xsd
    └── metadata.xml
    └── metadata.xsd
```

Kuvio 9: SIARD-tiedoston laajennettu header-kansio.

Tietovarastot konvertoitiin XML-muotoon tarkoitusta varten kehitetyn *DWXML (Data Warehouse Extensible Markup Language)* -kuvauskielen avulla. DWXML-dokumentti ja sitä vastaava XML-skeema liitettiin SIARD-pakettiin *DBPreserve Suite* -työkalun avulla. Tietokannan muuntamiseksi SIARD-muotoon käytti sovellus puolestaan SIARD Suite -työkalua (ks. luku 3.1.2). (Aldeias ym. 2011, 121–124.) Tietokannan säilytysmuotoon saattaminen DBPreserven avulla sisältää kaksivaiheisen konversion (ks. kuvio 10). Tietokannan relaatiomalli muunnetaan ensin moniulotteiseen malliin. Sitten moniulotteiseen malliin muunnettu tietokanta konvertoidaan XML-pohjaiseen säilytysformaattiin. (Aldeias ym. 2011, 116.)

Kuvio 10: Tietokannan muuntaminen DBPreserven säilytysmuotoon.

DBPreserve Suitea testattiin yksittäisessä tapaustutkimuksessa. Tutkimuksessa todettiin DWXML hyödylliseksi menetelmäksi tietovarastojen kuvailuun ja säilyttämiseen pitkällä aikavälillä (Aldeias 2011, 78).

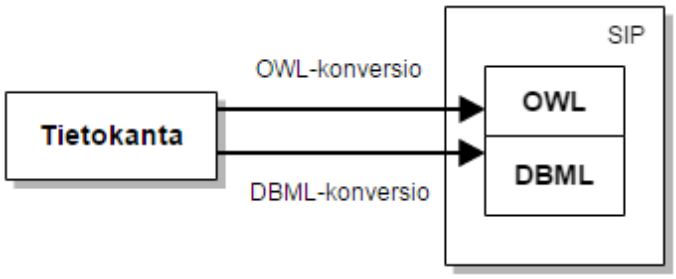
3.3 Ontologiat relaatiotietokantojen säilytyksessä

Freitas (2012) on väitöksessään *Relational Databases Digital Preservation* arvostellut XML-pohjaisia relaatiotietokantojen säilytysformaatteja siitä, että ne tyytyvät kuvaamaan tietokannan tietosisällön, mutta tietokannan semantiikkaa, eli tietokannan sisältämän tiedon merkitystä käsitteellisellä tasolla, ei säilytetä. Freitas esittää ratkaisuksi *ontologioiden* (engl. *ontology*) muodostamista tietokannan sisällöstä. Ontologian avulla voidaan tietokannasta muodostaa korkeamman tason käsitteellinen malli, joka liitetään tietokannan säilytystiedostoon. Ontologiat soveltuvat tarkoitukseen hyvin, sillä niiden

avulla voidaan kuvata kahden toisiinsa liittyvän tieto-objektin välisiä suhteita. Ontologian tallentaminen mahdollistaa tietokannan sisällön koneellisen tulkitsemisen, sekä uudenlaisen tavan tarkastella arkistoidun tietokannan sisältöä ontologiaselaimen avulla, jolloin käyttäjän ei tarvitse osata SQL-kyselykieltä. (Freitas 2012, 89, 133–134.)

Tutkimuksen tarkoituksena oli kehittää pitkäaikaissäilytysjärjestelmä, jossa relaatiotietokannan semanttinen esitysmuoto sisältyisi tietokannan siirtopakettiin XML-muodon ohella. Tietokannan rakenteen ja tietosisällön kuvaamiseen valikoitui DBML-formaatti yksinkertaisuutensa vuoksi (Freitas 2012, 61). Semanttiseen kuvailuun tutkimuksessa käytettiin W3C:n kehittämää *OWL (Web Ontology Language)* -kuvailukieltä, joka perustuu W3C:n standardoimaan *RDF (Resource Description Framework)* -metatietomalliin. Tietokannan siirtopaketti sisälsi täten kaksi erillistä esitysmuotoa tietokannasta: käsitteellisen kuvauksen OWL-muodossa, sekä rakenteen ja tietosisällön kuvauksen DBML-muodossa.

Tutkimuksessa kehitettiin prototyyppi OAIS-viitemalliin mukaisesta pitkäaikaissäilytysjärjestelmästä nimeltä *Framework for Relational DataBase Preservation* (jatkossa: FrepDB). Sovelluksen avulla tietokanta voitiin konvertoida DBML- ja OWL-muotoihin ja yhdistää nämä siirtopaketiksi (ks. kuvio 11). Tämän jälkeen siirtopaketti siirrettiin säilytysjärjestelmään, josta se voitiin tarjota edelleen kuluttajille. Järjestelmä tarjosi kaksi eri vaihtoehtoa arkistoidun tietokannan tarkastelemiseksi: tietokannan ontologian tarkastelun ontologiaselaimen avulla, sekä tietokannan palauttamisen toiminnassa olevaan tietokannanhallintajärjestelmään, jonka jälkeen siihen voitiin tehdä kyselyjä. (Freitas 2012, 89.)



Kuvio 11: Tietokannan muuntaminen FrepDB-järjestelmän siirtopaketiksi.

Ohjelmistoa testattiin tapaustutkimuksen avulla toiminnassa oleville tietokannoille. Testauksen pääpaino oli ontologian muodostamisessa sekä siirtopaketin kokoamisessa ja talteenotossa. OWL-dokumentin sisältöä verrattiin alkuperäisen tietokannan rakenteeseen ja sisältöön, sekä DBML-dokumentin sisältöön yhdenmukaisuuden varmistamiseksi. (Freitas 2012, 112.) Tutkimuksessa todettiin, että miljoonia tietueita sisältävien tietokantojen konvertoiminen tulee todennäköisesti olemaan hyvin hidasta, ja tuottaa hyvin suuria XML-tiedostoja, joiden käsitteleminen vaatii tehokkaan laitteiston. Ontologiamuunnoksen tuomia etuja tiedon tarkastelemiselle, käsittelylle ja koneelliselle tulkinnalle pidettiin kuitenkin hyödyllisinä ominaisuuksina tulevaisuuden digitaalisten säilytysjärjestelmien kehittämistä ajatellen (Freitas 2012, 134).

3.4 XML-konversio tieteellisen datan arkistoinnissa

Tietokantajärjestelmät on usein toteutettu siten, että päivitettäessä tietoa uusi arvo kirjoitetaan vanhan arvon päälle, eikä tieto muutoksesta tallennu mihinkään. Tällöin menetetään tietokannan muutoshistoria, ja arkistoon on mahdollista saada vain tuoreimman ajankohdan tilanne. Müllerin ym. mukaan tämä voi johtaa tieteellisenä näyttönä toimivan datan häviämiseen, jolloin löydösten perusteet eivät ole enää todennettavissa (Müller ym. 2008, 1295). Monessa tapauksessa tietokantaa käyttävä tutkija ei myöskään ole kiinnostunut yksittäisestä tietokannan otoksesta, vaan pitkälle menneisyyteen ulottuvasta pitkittäistutkimuksesta. SIARD:in ja RODA:n kaltaisten työkalujen avulla on tietokantoja mahdollista arkistoida vain yksi otos kerrallaan, ja eri otosten vertailemiseksi tulee kaikki tarkasteltavat otokset palauttaa tietokantajärjestelmään yksitellen, mikä voi olla työläs prosessi, jos otoksia tarvitaan pitkältä ajanjaksolta.

Edinburghin yliopistolla kehitetty XArch on pitkittäistutkimuksen mahdollistava ohjelmisto tieteellisen datan arkistointiin. XArchin avulla kaikki tietokannan eri versiot voidaan säilyttää yhdessä ja samassa XML-dokumentissa aikaleimoilla eroteltuina. Eri ajankohtina otetut otokset säilytetään samassa säilytystiedostossa; ainoastaan aiemmasta versiosta muuttunut tieto liitetään säilytystiedostoon. Jokaisella tietokannan elementillä on oma aikaleimansa, jonka avulla eri versiot erotetaan toisistaan. Kaikki tietokannan eri versiot säilytetään samassa XML-dokumentissa rinnakkain, mikä mahdollistaa myös ajallisten kyselyiden suorittamisen säilytettyyn tietokantaan ilman, että lukuisia otoksia on tarvetta palauttaa erikseen tietokannanhallintajärjestelmään. XArchin XML-muotoa

on havainnollistettu kuviossa 12. Eri vuosina tallennetut arvot on eroteltu toisistaan *T*-elementin avulla.

```
<OSASTO NIMI='Markkinointi'>
  <TYONTEKIJA ID='1'>
    <ETUNIMI>Miina</ETUNIMI>
    <SUKUNIMI>Matikainen</SUKUNIMI>
    <PALKKA>
      <T t='2010-2012'>30,000</T>
      <T t='2013-2015'>40,000</T>
    </PALKKA>
  </TYONTEKIJA>
...
</OSASTO>
```

Kuvio 12: XArchin XML-muotoa yksinkertaistettuna.

XArch on pitkäaikaissäilytysjärjestelmä, jonka avulla arkistoituja tietokantoja voidaan luoda ja hallita, sekä uusia versioita liittää olemassa oleviin arkistoihin. Aineistoon voidaan suorittaa kyselyjä erityisen XAQL-kyselykielen avulla (Müller ym. 2008, 1296). XAQL-kyselykielen avulla voidaan tarkastella joko tietokannan yksittäistä otosta, tai koko tietokannan muutoshistoriaa tietyllä ajanjaksolla. Myös yksittäisten tieto-objektien muutoshistoriaa on mahdollista tarkastella.

Tietokannan muuntamiseksi XML-kielelle XArch käyttää *PRATA (Publishing Relational Data Using Attribute Translation Grammars)* -nimistä järjestelmää, joka hakee tietokannan sisältämän datan tietokantajärjestelmästä, ja luo sen pohjalta XML-dokumentin (Benedikt ym. 2002). Relaatiotietokantojen lisäksi XArchin avulla on mahdollista käsitellä muun muassa CIA World Factbook -tietokannan, Gene Ontology -tietokannan, sekä Met Office -sääpalvelun tuottamia XML-tiedostoja.

3.5 Nykymenetelmien heikkoudet

SIARD:in ja DBML:n kaltaiset säilytysmenetelmät, joissa tietokanta muunnetaan XML-muotoon, ovat tällä hetkellä vallitseva käytäntö relaatiotietokantojen pitkäaikaissäilytyksessä. Käytännössä tietokantojen arkistointiprosessi on muuttunut kuitenkin hyvin vähän sitten 1980-luvun. Yleensä tietokannan arkistointiin sisältyy kolme vaihetta:

- Tietokannasta tallennetaan otos.
- Otos muunnetaan avoimeen säilytysformaattiin, ja mahdollisimman paljon tietokannan rakenteesta ja tietomallista pyritään säilyttämään alkuperäisen kaltaisena.
- Kun tietokannan sisältöä tarvitaan, palautetaan otos toiminnassa olevaan tietokannanhallintajärjestelmään kyselyjen tekemiseksi.

  (Delve ym. 2014.)

Tämä lähestymistapa on yleisesti ja laajalti hyväksytty, ja se onkin riittävä ratkaisu useimpien tahojen, kuten julkishallinnon ja tieteellisten arkistojen, tarpeisiin. Menetelmä on katsottu edulliseksi myös tietokannan autenttisuuden säilymisen kannalta. Tähän lähestymistapaan liittyy kuitenkin ongelmia tietokannan sisältämien tietojen käytettävyyden ja saavutettavuuden kannalta. Päästäkseen käsiksi tietokannan tietoihin joutuu käyttäjä käymään läpi seuraavat vaiheet:

- Selvitetään, missä tietokantaotos sijaitsee.
- Palautetaan se toiminnassa olevaan tietokannanhallintajärjestelmään.
- Suoritetaan tarvittavat kyselyt halutun tiedon löytämiseksi.

  (Delve ym. 2014.)

Yleensä käyttäjät hakevat hyvin spesifistä tietoa tietyltä aikaväliltä. Menetelmän heikkoutena onkin se, että esimerkiksi tietystä rakennuksesta tietoa etsivä henkilö ei voi hakea tietoa rakennuksen osoitteen perusteella, vaan hänen täytyy osata paikantaa juuri se tietokannan otos, joka sisältää yksityiskohtia juuri kyseisen alueen rakennuksista haluttuna ajankohtana. Vasta tämän jälkeen hän voi palauttaa kyseisen tietokannan tietokannanhallintajärjestelmään. Jos käyttäjä tarvitsee tietoa pidemmältä aikaväliltä, saattaa hän joutua palauttamaan useita otoksia tietokannasta tietokannanhallintajärjestelmään. Käyttäjän tulee lisäksi osata käyttää tietokannanhallintajärjestelmää, sekä hallita SQL-kyselykieltä, jotta voi löytää tietokannan otoksesta tai otosten joukosta juuri sen tiedon, jota tarvitsee. Tämän kaltainen toteutus säilytetyn tietokannan tietosisältöön käsiksi pääsemiseksi on työläs, hankala ja vaatii käyttäjiltä paljon teknistä osaamista. (ks. Delve ym. 2014.) Eräs tämän tutkielman tavoitteista on kartoittaa vaihtoehtoisia toteutustapoja säilytysmenetelmille, arvioida niitä, ja sitä kautta etsiä ratkaisuja näihin ongelmiin.

4 TUTKIMUSASETELMA JA -MENETELMÄT

Tässä luvussa esitellään ensin tutkimuskysymykset ja tutkimuksen tavoitteet yksityiskohtaisesti. Tämän jälkeen esitellään tutkimuksessa käytetyt tiedonhankinta- ja tutkimusmenetelmät, sekä aineisto, johon tutkimus pohjautuu.

4.1 Tutkimusongelman määrittely

Tietokannan konversio XML-muotoon on muodostunut yleisimmin käytetyksi relaatiotietokantojen pitkäaikaissäilytysmenetelmäksi. Eri tutkimusprojektit ovat kuitenkin päätyneet saman strategian käyttöön erilaisista lähtökohdista, ja toteutukset poikkeavat toisistaan. Tämän tutkimuksen ensisijaisena tavoitteena on kartoittaa, millaisia XML-konversioon pohjautuvia ratkaisuja relaatiotietokantojen pitkäaikaissäilytykseen tähän mennessä on kehitetty, ja tutkia, kuinka ne eroavat toisistaan. Tutkimuksessa tarkastellaan erityisesti sitä, kuinka eri säilytysmenetelmien tekninen toteutus ja kyky säilyttää alkuperäisen tietokannan ominaisuudet eroavat toisistaan. Säilytysmenetelmien vertailun tukena käytetään luvussa 2.4 esiteltyjä säilytyksen kannalta keskeisiä relaatiotietokannan ominaisuuksia, jotka on määritelty E-ARK-projektin yhteydessä (Ferreira 2016, 18–19). Vertailun pohjaksi valittiin seuraavat keskeiset ominaisuudet:

- Tietokannan sisältämä data.
- Relaatiorakenne, joka sisältää tietokannan skeemat, taulut, avaimet, sekä attribuutit ja niiden tietotyypit.
- Tietokannan käyttäytymiseen liittyvät tiedot, joihin kuuluvat tietokannan käyttäjät, roolit ja käyttöoikeudet, näkymät, herättimet, tallennetut proseduurit ja funktiot, sekä tietokannan rajoitteet.

Keskeisiksi ominaisuuksiksi voidaan E-ARK-projektin määritelmän mukaisesti katsoa myös tietokannasta tallennetut metatiedot, jotka kuvailevat tietokannan eri ominaisuuksia, kuten sen kenttiä ja tietotyyppejä, sekä tietokantasovelluksen dokumentaatio, josta käy ilmi, millaisessa vuorovaikutuksessa tietokantasovellus ja tietokanta ovat olleet (ks. Ferreira 2016, 19–20). Koska tutkielma painottuu säilytysmenetelmien teknisen toteutuksen tarkasteluun, ja kirjallisuuden osalta metatietoon ja dokumentaatioon liittyvät osa-alueet koettiin puutteellisiksi, jäävät nämä ulottuvuudet tarkastelun ulkopuolelle. Eri

menetelmien nopeutta ja tehokkuutta tietokannan konvertoimisessa ei tässä tutkimuksessa myöskään pyritä vertailemaan, sillä se olisi mahdotonta ilman yksityiskohtaisia tapaustutkimuksia ja yhteismitallisia mittausvälineitä.

Teknisen toteutuksen lisäksi tutkielmassa tarkastellaan sitä, kuinka tietokannan haku- ja prosessointimahdollisuudet säilyvät arkistointiprosessissa. Tietokannan palauttaminen toiminnassa olevaan tietokannanhallintajärjestelmään on muodostunut tyypilliseksi ratkaisuksi säilytetyn tietokannan tietosisällön käsittelemiseen ja tarkastelemiseen. Tutkimuksen tarkoituksena on kartoittaa, millaisia vaihtoehtoisia menetelmiä tarkoitukseen on kehitetty.

Tutkimuskysymykset ovat:

- Kuinka kattavasti kirjallisuuden pohjalta tunnistetut, XML-konversioon pohjautuvat säilytysmenetelmät säilyttävät relaatiotietokantojen keskeiset ominaisuudet?

- Millaisia menetelmiä säilytettyjen relaatiotietokantojen tietosisällön tarkastelemiseksi ja käsittelemiseksi on kehitetty?

Tutkimuksen ensisijaisena tavoitteena on koota yhteen olemassa oleva tieto relaatiotietokantojen pitkäaikaissäilytyksestä XML-konversion avulla. Tämän kaltaista kokoavaa ja vertailevaa tutkimusta ei ole aiheesta aikaisemmin Suomessa tehty. Tutkimuksen tarkoitus on antaa taustatietoa tietokantojen pitkäaikaissäilytyksestä arkistonmuodostajille ja arkistonhoitajille, sekä tuoda aihepiiriä informaatiotutkimuksen ja tietojenkäsittelytieteen tutkijoiden tietouteen. Pyrkimyksenä on tuoda alan tutkijoille ajankohtaista tietoa tutkimuksen tämän hetkisestä tilasta sekä mahdollisesta aiheeseen liittyvän lisätutkimuksen tarpeesta.

4.2 Tutkimusmenetelmän valinta

Tutkimusmenetelmäksi valikoitui systemaattinen kirjallisuuskatsaus, joka on tehokas menetelmä jo olemassa olevan tutkimustiedon syventämiseksi. Tutkimuskirjallisuuteen perustuva kirjallisuuskatsaus on systemaattinen, täsmällinen ja toistettavissa oleva menetelmä, jolla tunnistetaan, arvioidaan ja tiivistetään jokin olemassa oleva ja julkaistu tutkimusaineisto (Fink 2013, 3). Tutkimustulosten analysoinnin ja synteesin avulla voi-

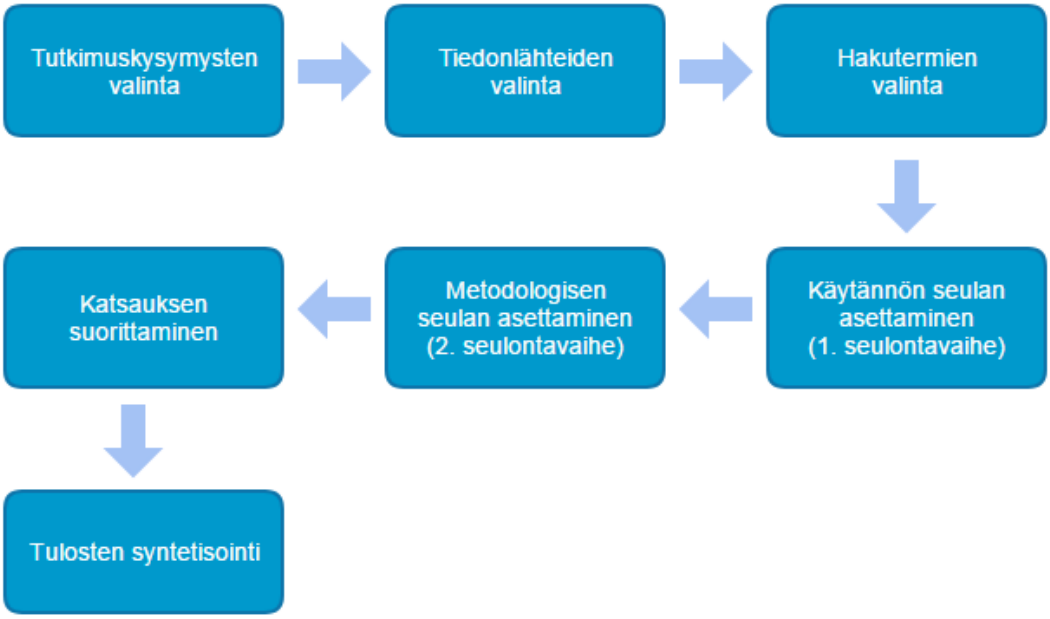
daan tuottaa kokonaan uusia näkökulmia tutkimusaineistoon. Kirjallisuuskatsauksen tavoitteena ei ole vain arvioida teoriaa, vaan kehittää olemassa olevaa teoriaa ja rakentaa sen pohjalta uutta teoriaa. Systemaattista kirjallisuuskatsausta voidaan hyödyntää hypoteesien testaamiseen, tutkimusten tulosten esittämiseen tiiviissä muodossa, sekä niiden johdonmukaisuuden arvioimiseen. Sen avulla voidaan havaita aikaisemmassa tutkimuksessa esiintyviä puutteita ja tuoda esiin uusia tutkimustarpeita. (Salminen 2011, 3–5, 9.) Tavoitteena ei ole tuottaa pelkästään yhteenvetoa tutkimusaineistosta, vaan tulkita tuloksia, selittää tutkimusaineistossa ilmeneviä eroavaisuuksia tai samankaltaisuuksia, ja sitä kautta tuottaa uutta tietoa, jota ei yksittäisiä tutkimusartikkeleita lukemalla olisi mahdollista havaita (Aveyard 2010, 124).

Systemaattisessa kirjallisuuskatsauksessa huomiota kiinnitetään erityisesti käytettyjen lähteiden keskinäiseen yhteyteen, sekä tekniikkaan, jolla siteeratut tulokset on hankittu (Salminen 2011, 4). Aineisto seulotaan huolellisesti tarkoin valittuja seulontakriteerejä noudattaen, mikä luo tutkimukselle uskottavuutta, ja auttaa varmistumaan siitä, että lähdeaineistot ovat keskenään loogisia. (Salminen 2011, 11). Huolellisella tiedonhakustrategialla varmistutaan siitä, että kaikki olennainen kirjallisuus päätyy mukaan katsaukseen (Aveyard 2010, 14).

Fink (2013, 3–5) on esittänyt seitsenvaiheisen mallin systemaattisen kirjallisuuskatsauksen tekemiseksi (ks. kuvio 13). Prosessi aloitetaan asettamalla tutkimuskysymys. Tämän jälkeen valitaan käytettävä kirjallisuus ja tietokannat. Seuraavassa vaiheessa valitaan hakutermit, joiden avulla hakua pyritään rajaamaan siten, että materiaali parhaiten vastaisi tutkimuskysymykseen. Neljännessä vaiheessa asetetaan käytännön seulontakriteerit, joiden avulla hakutuloksia karsitaan. Materiaalia voidaan seuloa muun muassa sisällön, kielen ja julkaisuajankohdan perusteella. Viidennessä vaiheessa asetetaan metodologiset seulontakriteerit, joissa painotus kohdistuu tutkimusten tekotapoihin, tuloksiin ja johtopäätöksiin. Tarkoituksena on valikoida katsaukseen vain laadukkain mahdollinen materiaali. Kuudennessa vaiheessa suoritetaan itse katsaus. (Fink 2013, 3–5.)

Prosessin seitsemännessä eli viimeisessä vaiheessa tuotetaan synteesi tuloksista. Synteesin tarkoituksena on raportoida tämänhetkinen tieto, selittää löydökset ja osoittaa jatkotutkimuksen tarpeet. Tutkimuksen laadun tarkkailu on olennainen osa synteesin tekemistä. (Fink 2013, 4–5.) Baumeisterin ja Learyn (1997; tässä Salminen 2011, 10) mu-

kaan synteesi on systemaattisessa kirjallisuuskatsauksessa kriittinen vaihe, jossa monet epäonnistuvat. Riskinä on, että tutkimusten integrointi jää liian pinnalliseksi kuvailuksi.



Kuvio 13: Finkin systemaattisen kirjallisuuskatsauksen malli yksinkertaistettuna.

Tässä tutkielmassa sovelletaan systemaattisen kirjallisuuskatsauksen piirteitä erityisesti tiedonhankinnassa, seulontakriteerien määrittämisessä ja aineiston valinnassa.

4.3 Tiedonhaku ja aineiston valinta

Kirjallisuuskatsausta varten suoritettiin laaja tiedonhaku kansainvälisiin informaatiotutkimuksen ja tietojenkäsittelytieteen tietokantoihin maalis-huhtikuussa 2017. Suomessa relaatiotietokantojen pitkäaikaissäilytyksen tutkimus on ollut vähäistä, joten haku keskitettiin kansainvälisiin tietokantoihin.

Tutkimusaineiston rajauksessa ja valinnassa käytettiin seuraavia kriteerejä:

- *Artikkeli käsittelee relaatiotietokantojen pitkäaikaissäilytystä XML-konversion avulla.* Oletuksena on, että valtaviin tietokantoihin suoritetut haut palauttavat paljon tutkimuksen kannalta epärelevantteja julkaisuja, jotka on suodatettava pois.
- *Aineisto on kirjoitettu englannin tai suomen kielellä.* Kielet on valittu tutkielman tekijän kielitaidon perusteella.

- *Aineisto on julkaistu vuoden 2000 jälkeen.* Aikarajaus pyrittiin asettamaan siten, että kaikki olennainen kirjallisuus päätyy katsaukseen mukaan.
- *Ulkomaisen lehtiartikkelin tulee olla julkaistu, läpäissyt vertaisarvioinnin ja saatavilla kokonaisuudessaan.* Tämän kriteerin avulla pyritään varmistumaan tutkimusaineiston laadukkuudesta.

Tutkimuksen ulkopuolelle suljettiin artikkelit, jotka oli julkaistu ennen vuotta 2000; artikkelit, joita ei ollut julkaistu lainkaan; sekä artikkelit, joita ei ollut saatavilla kokonaisuudessaan. Myös muita tietokantatyyppejä kuin relaatiotietokantoja käsittelevät artikkelit, sekä muihin kuin XML-konversioon pohjautuvia säilytysmenetelmiä käsittelevät artikkelit suljettiin tutkimuksen ulkopuolelle.

Aineiston kerääminen aloitettiin määrittämällä hakutermit tutkimuskysymysten pohjalta. Tiedonhaku suoritettiin seuraavilla englanninkielisillä hakutermeillä ja niiden yhdistelmillä:

- *relational database*
- *long-term preservation*
- *archive*

Hakua täydennettiin termeillä:

- *xml*
- *conversion*
- *migration*

Tiedonhaku suoritettiin seuraavissa kansainvälisissä tietokannoissa:

1. *Academic Search Premier* – Monitieteinen kokotekstitietokanta, joka sisältää yli 3 900 tieteellisesti arvioitua lehteä eri aloilta.
2. *ACM Digital Library* – Sisältää tietojenkäsittelytieteen artikkeleita sekä artikkeliviitteitä lehdistä ja konferenssijulkaisuista.
3. *Emerald* – Monitieteinen kokotekstitietokanta, joka sisältää vertaisarvioituja lehtiartikkeleita informaatiotutkimuksen alalta.

4. *IEEE Xplore Digital Library* – Sisältää tietotekniikan, elektroniikan ja sähkötekniikan alojen verkkolehtiä ja konferenssijulkaisuja.
5. *Library & Information Science Abstracts (LISA)* – Kirjastotiedettä ja informaatiotutkimusta käsitteleviä kirjallisuusviitteitä.
6. *Library, Information Science & Technology Abstracts (LISTA)* – Sisältää informaatiotutkimukseen liittyviä kirjallisuusviitteitä abstrakteineen.
7. *Science Direct* – Monitieteinen kokotekstitietokanta, joka sisältää yli 2 000 lehden artikkelit muun muassa luonnontieteiden, lääketieteen ja yhteiskuntatieteen aloilta.
8. *SpringerLink* – Sisältää yli 1 800 lehteä lääketieteen ja luonnontieteiden, kuten tietojenkäsittelytieteen, aloilta.

Tietokannat on valittu aiherelevanssin mukaisesti tietojenkäsittelytieteen ja informaatiotutkimuksen aloilta. Hakua laajennettiin lisäksi *Google Scholar* -hakupalvelun avulla mahdollisimman kattavan kuvan saamiseksi.

Tiedonhaussa noudatettiin Finkin (2013, 5) esittämää kaksivaiheista seulontaprosessia. Tiedonhaun ensimmäisessä vaiheessa asetettiin käytännön seula, jolloin hakua rajattiin kielen ja julkaisuajankohdan perusteella (ks. Fink 2013, 5). Tämän jälkeen hakutuloksista seulottiin tutkimuksen kannalta relevantit artikkelit lukemalla niiden otsikot ja tiivistelmät. Hakutulosten suuren määrän vuoksi tulosten tarkastelu rajattiin 200 ensimmäiseen hakutulokseen kussakin hakupalvelussa hakupalvelun oman, sisäisen relevanssin mukaan järjestettynä. Otsikoiden ja tiivistelmien perusteella valikoitui kansainvälisistä tietokannoista yhteensä 26 artikkelia toiseen seulontavaiheeseen.

Tiedonhaun toisessa vaiheessa asetettiin metodologinen seula (ks. Fink 2013, 5). Toisen seulontavaiheen tarkoituksena oli rajata aineistoa siten, että se vastaa mahdollisimman hyvin tutkimusongelmaan. Huomio kiinnitettiin tarkasteltavien tutkimusten tuloksiin ja niiden tieteellisen laadun arviointiin. Tarkastelun kohteena olivat muun muassa tutkimuksessa käytetyt tutkimusmenetelmät ja niiden soveltuvuus, sekä tutkimusaihe ja sen relevanssi. Näitä seulontakriteerejä sovellettiin perehtymällä artikkeleiden ja tutkimusten kokoteksteihin. Tässä tiedonhaun vaiheessa oli tarkoituksena pudottaa pois ”kaikki sellainen, joka on tutkimuskysymyksen kannalta merkityksetöntä tai ei tuo tutkimuksel-

le lisäarvoa" (Salminen 2011, 10). Lopulliselta aineistolistalta poistettiin artikkelit, joissa käsiteltiin relaatiotietokantojen pitkäaikaissäilytystä tai XML-konversiota yleisellä tasolla, sekä artikkelit, joissa tutkimuksen painopiste oli muualla kuin XML-konversioon pohjautuvan säilytysmenetelmän tutkimuksessa. Lopulliseen aineistoon päätyi seulonnan tuloksena 12 artikkelia. Seulonnan vaiheet on kuvattu taulukossa 2.

Tietokanta	Hakutuloksia	1. seulonta	2. seulonta
Academic Search Premier	68	1	0
ACM	8 841	5	3
Emerald	545	1	0
IEEE Xplore	112	2	0
Library & Information Science Abstracts (LISA)	637	1	0
Library, Information Science & Technology Abstracts (LISTA)	41	0	0
Science Direct	3970	0	0
SpringerLink	1 484	3	2
Google Scholar	17 200	13	7
Yhteensä	**32 698**	**26**	**12**

Taulukko 2: Kansainvälisissä tietokannoissa suoritetun aineistonhaun tulokset.

Tiedonhakua täydennettiin tutkimusprosessin aikana manuaalisesti tutkimalla jo löydettyjen artikkelien lähdeviitteitä. Hakua laajennettiin yhä laajemmalle alueelle seuraamalla viittausten ketjuja. Näin voitiin varmistua siitä, että kaikki olennainen kirjallisuus päätyi mukaan tutkimukseen (ks. mm. Choo ym. 2000). Manuaalisen täydennyshaun avulla löytyi yhteensä 31 artikkelia, joista luettiin kokotekstit. Lopulliseen aineistoon näistä valikoitui kaksi artikkelia.

4.4 Aineisto

Tutkielmassa kartoitettiin olemassa olevien XML-pohjaisten relaatiotietokantojen säilytysmenetelmien kirjo. Tavoitteena oli tuottaa kattava kuva alan kirjallisuudesta, ja tarkastella sen avulla eri menetelmien kykyä säilyttää tietokantojen keskeiset ominaisuudet

sekä eri menetelmien mahdollisuuksia säilytetyn tietokannan tietosisällön tarkastelemiseksi. Katsauksessa käytiin läpi kaiken kaikkiaan 57 artikkelia ja tutkimusta. Tutkielman lopulliseen aineistoon valikoituivat tiukkoja seulontakriteerejä noudattaen ne lähteet, jotka parhaiten vastasivat tutkimusongelmaan. Lopullinen aineisto sisältää konferenssikirjoituksia (11 kpl), tieteellisiä artikkeleita (2 kpl), sekä väitöskirjatutkimuksia (1 kpl). Artikkelit sekä tutkimukset on julkaistu vuosina 2004-2016. Aineisto on esitelty taulukossa 3.

Artikkelin nimi	Julk.v.	Julkaisu	Kirjoittajat
Providing authentic long-term archival access to complex relational data	2004	Proceedings PV-2004: Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data. 5.-7. lokakuuta, 2004. Frascati, Italia.	Heuscher, S. Järmann, S. Keller-Marxer, P. Möhle, F.
Relational database preservation through XML modeling	2007	Proceedings of the International Workshop of Overlapping Structures - Extreme Markup Languages. 7.-10. elokuuta, 2007. Montreal, Kanada.	Ramalho, J. C. Ferreira, M. Faria, L. Castro, R.
RODA -Repository of Authentic Digital Objects	2007	Proceedings of the International Workshop on Database Preservation (PresDB'07). 23. maaliskuuta, 2007. Edinburgh, Skotlanti.	Faria, L. Castro, R.
RODA and CRiB a service-oriented digital repository	2008	Proceedings of the International Conference on Preservation of Digital Objects (iPRES2008). 29.-30. syyskuuta, 2008. Lontoo, Englanti.	Ramalho, J. C. Ferreira, M. Faria, L. Castro, R. Barbedo, F.
XArch: archiving scientific and reference data	2008	Proceedings of the 2008 ACM SIGMOD international conference on Management of data (SIGMOD '08). 9.-12. kesäkuuta, 2008. Vancouver, Kanada.	Müller, H. Buneman, P. Koltsidas, I.
Curating the CIA World Factbook	2009	The International Journal of Digital Curation, vol. 3, issue 4, 2009.	Buneman, P. Müller, H. Rusbridge, C.
MIXED: Repository of Durable File Format Conversions	2009	Proceedings of the Sixth International Conference on the Preservation of Digital Objects (iPRES2009). 5.-6. lokakuuta, 2009. San Francisco, Kalifornia, Yhdysvallat.	Van Horik, R. Roorda, D.
RODA. A service-oriented repository to preserve authentic digital objects	2009	Proceedings of the 4th International Conference on Open Repositories. 18.-21. toukokuuta, 2009. Atlanta, Georgia, Yhdysvallat.	Barbedo, F. Castro, R. Corujo, L. Faria, L. Ferreira, M. Henriques, C. Ramalho, J.

Artikkelin nimi	Julk.v.	Julkaisu	Kirjoittajat
DWXML - A Preservation Format for Data Warehouses!	2011	Proceedings of the Ninth National Conference on XML, Associated Technologies and Applications (XATA 2011). 1.-2. kesäkuuta, 2011. Vila do Conde, Portugali.	Aldeias, C. David, G. Ribeiro, C.
Migration to Intermediate XML for Electronic Data (MIXED): Repository of Durable File Format Conversions	2011	The International Journal of Digital Curation, vol. 2, issue 6, 2011.	Van Horik, R. Roorda, D.
Relational databases digital preservation	2012	Väitöskirja. Minhon yliopisto, Portugali	Freitas, R.
SIARD Archive Browser	2012	Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (toim.) Theory and Practice of Digital Libraries. TPDL 2012. Lecture Notes in Computer Science, vol 7489. Springer, Berlin, Heidelberg.	Rahman, A. U. David, G. Ribeiro, C.
Database Preservation Toolkit: a flexible tool to normalize and give access to databases	2014	DLM Forum - 7th triennial conference. Making the Information Governance Landscape in Europe. 12.-14. marraskuuta, 2014. Lissabon, Portugali.	Ramalho, J. Faria, L. Silva, H. Coutada, M.
Database Preservation Toolkit	2016	Proceedings of the 13th International Conference on Digital Preservation (iPRES2016). 3.-6. lokakuuta, 2016. Bern, Sveitsi.	Ferreira, B. Faria, L. Ramalho, J. Ferreira, M.

Taulukko 3: Systemaattiseen kirjallisuuskatsaukseen valitut artikkelit.

4.5 Aineiston analyysi

Aineiston analyysi, tulkinta ja johtopäätösten teko ovat olennaisia vaiheita tutkimuksen tekemisessä. Vasta analyysivaiheessa selviää, millaisia vastauksia tutkimusongelmiin saadaan. Laadulliselle tutkimukselle on tyypillistä, että aineistoa kerätään monissa vaiheissa ja rinnakkaisesti eri menetelmin, jolloin aineiston keruu ja analysointi limittyvät. (Hirsjärvi ym. 2007, 218–219.) Tutkimusongelman, käsitteiden ja määritteiden valinta on osa analysointiprosessia. Analyysimenetelmät kehitetään kerätyn aineiston pohjalta, ja sen hetkisten teknisten mahdollisuuksien mukaan. Tavoitteena on, että analyysimenetelmät palvelisivat tutkimuksen tarkoituksia parhaalla mahdollisella tavalla. (Grönfors 2011, 85.)

Laadullisessa tutkimuksessa analyysi ja synteesi yhdistyvät aineiston analysoinnissa. Analyyttisen prosessin avulla kerätty aineisto hajotetaan pienemmiksi, käsitteellisiksi osiksi, jotka kootaan uudelleen synteesin avulla tieteellisiksi johtopäätöksiksi. Analyy-

silla ei tarkoiteta tässä tutkimustekniikoihin liittyviä mekaanisia toimenpiteitä, vaan analyysi ja synteesi ovat järjellistä toimintaa, jossa tutkimusaineistoa tarkastellaan käsitteellisellä tasolla. (Grönfors 2011, 85.)

Saldana ym. (2014, 90, 95) korostavat lukemisen ja kirjoittamisen merkitystä aineiston analysoinnissa. Toistuvat lukukerrat tutustuttavat tutkijan läheisesti tutkimusaineistoonsa, sekä auttavat havaitsemaan tärkeitä yksityiskohtia, ja tekemään oivalluksia niiden pohjalta. Muistiinpanojen kirjoittaminen, merkintöjen tekeminen ja muu aineiston dokumentoiminen puolestaan auttaa omaksumaan tietoa, tunnistamaan aineistossa esiintyviä kaavamaisuuksia, löytämään yhteyksiä aineistojen välillä, sekä luomaan kokonaiskuvaa aineistosta. (Saldana ym. 2014, 90, 95.) Myös Hirsjärvi ym. (2007, 33) korostavat kirjoittamisen merkitystä tutkimuksen teossa. Kirjoittaminen aktivoi ajattelua ja sitoo tutkijan tutkittavaan asiaan. Näin kirjoittaminen ja ajattelu kietoutuvat tutkimusprosessissa toisiinsa. (Hirsjärvi ym. 2007, 33.)

Koska kyseessä on laadullinen tutkimus, aloitettiin tämän tutkimuksen aineiston analysoiminen heti, kun aineistoa oli ryhdytty keräämään ja käymään läpi. Aineistoa jäsennettiin samaan aikaan sen lukemisen yhteydessä muun muassa taulukoimalla, mikä auttoi löytämään aineistosta toistuvia seikkoja ja tunnistamaan keskeisiä käsitteitä. Aineistoa järjestettiin useiden toisiaan seuranneiden lukukertojen aikana, mikä auttoi ymmärtämään aineiston keskeistä sisältöä, löytämään yhteyksiä eri aineistojen välillä, ja tekemään johtopäätöksiä aineiston pohjalta. Myös kirjoittamisella oli merkittävä rooli analysointiprosessissa. Kirjoittaminen auttoi hahmottamaan aineiston keskeistä sisältöä, sekä luomaan vuoropuhelua aineiston ja tutkimuskysymysten välille, ja sitä kautta löytämään aineistosta vastauksia tutkimusongelmaan.

5 TULOKSET

Tutkielman tarkoituksena oli kartoittaa tällä hetkellä olemassa olevien, XML-konversioon pohjautuvien relaatiotietokantojen pitkäaikaissäilytysmenetelmien kirjo, sekä vertailla eri menetelmien kykyä säilyttää alkuperäisen tietokannan ominaisuudet. Systemaattisessa kirjallisuuskatsauksessa tarkasteltiin säilytysmenetelmien välisiä eroja luvussa 4.1 esiteltyjä keskeisiä ominaisuuksia soveltaen. Lisäksi tutkielmassa tarkasteltiin sitä, millaisia eri tapoja tietokannan tietosisällön käsittelemiseen ja tarkastelemiseen oli eri menetelmissä kehitetty. Eri menetelmien nopeutta ja tehokkuutta tietokannan konvertoimisessa ei tässä tutkimuksessa pyritty vertailemaan, sillä se olisi mahdotonta ilman yksityiskohtaisia tapaustutkimuksia ja yhteismitallisia mittausvälineitä.

Kirjallisuuskatsauksessa tunnistettiin neljä XML-konversioon pohjautuvaa relaatiotietokantojen säilytysformaattia:

- DBML (Database Markup Language)
- DBPreserve - *prototyyppi*
- SDFP (Standard Data Formats for Preservation)
- SIARD (Software Independent Archiving of Relational Databases)

Lisäksi tunnistettiin kaksi digitaalista pitkäaikaissäilytysjärjestelmää, joissa XML-konversio sekä siihen pohjautuva säilytysformaatti ovat integroitu järjestelmään:

- FrepDB (Framework for Relational Database Preservation) - *prototyyppi*
- XArch

SDFP-säilytysformaatin osalta kirjallisuus todettiin puutteelliseksi, minkä vuoksi se ei päätynyt mukaan lopulliseen vertailuun. On myös hyvä huomata, että DBPreserve- ja FrepDB-ohjelmistot olivat tutkielmaa tehtäessä vasta prototyyppiasteella, eikä käytännön sovelluksia niistä ollut vielä saatavilla.

5.1 Säilytysmenetelmien vertailua

Tutkituista säilytysmenetelmistä DBML, DBPreserve ja SIARD ovat XML-pohjaisia säilytysformaatteja, jotka tarvitsevat erillisiä työkaluja toimiakseen. Kullekin säilytysformaatille on kehitetty omat ohjelmistonsa, joiden avulla säilytettävä tietokanta tai sen otos haetaan tietokannanhallintajärjestelmästä ja konvertoidaan säilytysformaattiin. Tämän jälkeen säilytystiedostoon liitetään ohjelmiston avulla metatietoa, jonka jälkeen tiedosto on valmis siirrettäväksi PAS-järjestelmään.

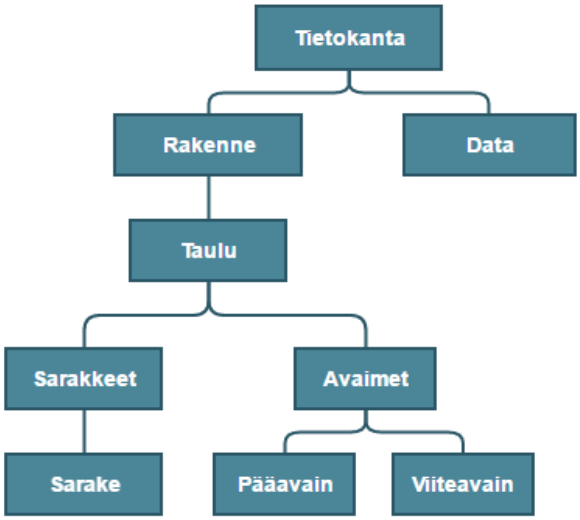
Tutkitut pitkäaikaissäilytysjärjestelmät – FrepDB ja XArch – puolestaan sisältävät sekä tarvittavat työkalut tietokannan konvertoimiseen, että PAS-arkiston toiminnot itse järjestelmään integroituina. Tietokannan XML-konversio suoritetaan osana järjestelmän toimintaa, eikä niiden käyttämiä säilytysformaatteja täten voi käyttää järjestelmästä erillään.

XArch poikkeaa muista tutkituista säilytysmenetelmistä. Se on digitaalinen pitkäaikaissäilytysjärjestelmä, jonka erikoisuutena on, että samasta tietokannasta eri ajankohtina tallennetut otokset voidaan yhdistää samaan säilytystiedostoon, jolloin säilyy myös tietokannan muutoshistoria. XArchilla on oma, muista menetelmistä poikkeava säilytysformaattinsa, jossa eri otoksiin liittyvä tieto on erotettu toisistaan aikaleimoilla. XArch eroaa muista menetelmistä myös siinä suhteessa, että se pakkaa säilytystiedoston, jotta se vie vähemmän tilaa. Tilansäästöä saavutetaan sekä eri otosten yhdistämisen, että pakkaamisen kautta, mikä voi tuoda kustannussäästöjä, kun on tarpeen arkistoida suuria tietokantoja.

Tutkimuksesta ilmeni, että XArch-järjestelmää lukuun ottamatta kaikki tarkastellut säilytysmenetelmät ovat OAIS-yhteensopivia. Vaikka SIARD ei ole erityisesti OAIS-yhteensopivuutta silmällä pitäen kehitetty, voidaan sitä käyttää OAIS-järjestelmän mukaisena siirtopakettina. XArch ei päivittyvän luonteensa vuoksi sovellu OAIS-yhteensopivassa järjestelmässä käytettäväksi. Koska useita otoksia samasta tietokannasta voidaan yhdistää samaan XML-tiedostoon, ei OAIS-viitemallissa kuvattua talteenottovaihetta voida näin ollen soveltaa XArch-järjestelmässä.

Kaikki tutkitut menetelmät perustuvat avoimeen lähdekoodiin ja avoimiin standardeihin, mikä mahdollistaa säilytysmenetelmien muokkaamisen organisaation omiin tarpeisiin. Tuettujen tietokantaohjelmistojen osalta säilytysmenetelmät eivät olennaisesti eroa toisistaan. Kaikki tutkitut menetelmät sisältävät tuen lähes kaikille tunnetuimmille tietokantaohjelmistoille.

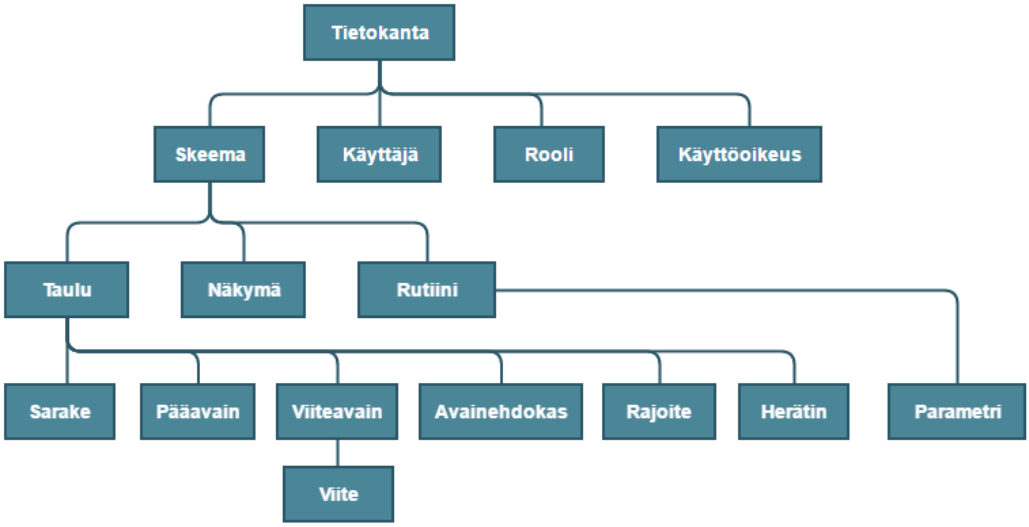
Kirjallisuuden perusteella oli havaittavissa selkeitä eroja eri säilytysmenetelmien toteutuksessa. Tutkituista säilytysformaateista yksinkertaisimmaksi osoittautui DBML. DBML sisältää tarvittavan metatiedon, sekä tietokannan rakenteen ja datan yhdessä ja samassa tiedostossa. Suuret *LOB (Large Object File)* -tiedostot tallennetaan säilytystiedostoon erikseen, ja niihin luodaan linkki DBML-dokumentista. DBML-dokumentin rakenne on kuvattu kuviossa 14.



Kuvio 14: DBML-dokumentin rakenne.

FrepDB:n säilytysformaatti perustuu DBML-formaattiin, eikä FrepDB:n DBML-toteutus poikkea normaalista DBML-muodosta millään tavalla. FrepDB:n säilytysformaattia on kuitenkin laajennettu OWL-ontologiadokumentilla, joka on linkitetty DBML-dokumenttiin, ja sisältää tietokannan käsitteellisen mallin.

SIARD on huomattavasti DBML:ää kehittyneempi tiedostomuoto. SIARD:ssa tieto on hajautettu säilytystiedoston sisällä siten, että metatiedot kuvataan omassa XML-tiedostossaan, ja jokainen tietokannan skeema ja taulu omassa XML-tiedostossaan. SIARD-formaatin rakennetta on havainnollistettu kuviossa 15.



Kuvio 15: SIARD-säilytystiedoston rakenne.

DBPreserve laajentaa SIARD-formaattia liittämällä SIARD-tiedostoon DWXML-dokumentin. Tämä tuottaa tietokannasta moniulotteisen mallin relaatiomalliin tallennetun tietokannan ohella.

XArch puolestaan tallentaa tietokannan sisällön yksinkertaisessa, hierarkkisessa XML-muodossa, jossa tietokannan alkuperäistä rakennetta tai attribuuttien tietotyyppejä ei ilmaista millään tavoin.

5.2 Keskeisten ominaisuuksien säilyminen

Tutkimuksessa tarkasteltiin sitä, kuinka kattavasti eri säilytysmenetelmät säilyttävät tietokannan keskeiset ominaisuudet arkistointiprosessissa. Tutkimuksessa tarkasteltavat relaatiotietokantojen ominaisuudet perustuivat E-ARK-projektissa määriteltyihin keskeisiin ominaisuuksiin, jotka on esitelty luvussa 4.1. Keskeisiksi ominaisuuksiksi katsottiin näin ollen tietokannan tietosisältö (data), relaatiorakenne (skeema, taulut, sarakkeet, avaimet), sekä tietokannan käyttäytymiseen liittyvät tiedot (käyttäjät, roolit, käyttöoikeudet, näkymät, herättimet, tallennetut proseduurit ja funktiot, sekä tietokannan rajoitteet).

5.2.1 Tietokannan data

Kaikki tutkitut menetelmät säilyttävät tietokannan datan, mikä oli odotettavissa.

5.2.2 Tietokannan relaatiorakenne

Lähes kaikki tutkitut menetelmät pyrkivät säilyttämään tietokannan relaatiorakenteen sellaisenaan, mikä oli odotettavissa (ks. taulukko 4). Kirjallisuuden perusteella voitiin kuitenkin päätellä, että XArch-järjestelmässä ei tietokannan skeemoja ja siten täydellistä relaatiorakennetta pyritä säilyttämään.

	DBML	DBPreserve	FrepDB	SIARD	XArch
Data	✓	✓	✓	✓	✓
Skeemat	✓	✓	✓	✓	✗
Taulut	✓	✓	✓	✓	✓
Sarakkeet	✓	✓	✓	✓	✓
Avaimet	✓	✓	✓	✓	✓

Taulukko 4: Tietokannan relaatiorakenteeseen liittyvät ominaisuudet.

5.2.3 Tietokannan käyttäytymiseen liittyvät tiedot

Tutkituista säilytysformaateista ainoastaan SIARD ja siihen perustuva DBPreserve pyrkivät säilyttämään tietokannan käyttäytymiseen liittyviä tietoja. SIARD-säilytysformaatti ja siten myös DBPreserve-säilytysformaatti mahdollistavat tallennettujen proseduurien ja funktioiden, herättimien, näkymien, käyttöoikeuksien ja käyttäjien, sekä tietokannan rajoitteiden säilyttämisen (ks. taulukko 5). Huomionarvoista kuitenkin on, että näissäkin säilytysmenetelmissä tietokannan käyttäytymiseen liittyvät tiedot ainoastaan dokumentoidaan säilytystiedostossa; niiden toiminnallisuus ei säily palautettaessa tietokanta toiminnassa olevaan tietokannanhallintajärjestelmään.

	DBML	DBPreserve	FrepDB	SIARD	XArch
Tallennetut prose-duurit	✗	✓	✗	✓	✗
Funktiot	✗	✓	✗	✓	✗
Herättimet	✗	✓	✗	✓	✗
Näkymät	✗	✓	✗	✓	✗
Käyttöoikeudet	✗	✓	✗	✓	✗
Käyttäjät	✗	✓	✗	✓	✗
Roolit	✗	✓	✗	✓	✗
Rajoitteet	✗	✓	✗	✓	✗

Taulukko 5: Tietokannan käyttäytymiseen liittyvät ominaisuudet.

5.3 Pääsy tietokannan tietosisältöön

Tyypillinen tapa toteuttaa pääsy säilytettyyn tietokantaan on palauttaa tietokanta toiminnassa olevaan tietokannanhallintajärjestelmään, jonka jälkeen siihen voidaan tehdä kyselyjä. Eräs tutkielman tavoitteista oli tutkia sitä, millaisia vaihtoehtoisia tapoja säilytetyn tietokannan tietosisällön tarkastelemiseksi ja käsittelemiseksi on kehitetty. Tutkimuksesta ilmeni, että eri säilytysmenetelmiä varten on kehitetty hyvin erilaisia, toisistaan poikkeavia ratkaisuja. Nämä on koottu taulukkoon 6.

Tutkituista menetelmistä DBPreserve tarjoaa mahdollisuuden tarkastella ja analysoida tietovarastoksi muunnettua tietokantaa XML:n valmiiden XPath- ja XQuery-työkalujen avulla. Moniulotteisen tietovarastomallin ansiosta tietokannan rakenne yksinkertaistuu, jolloin tietokannan sisältöä on helpompi ymmärtää alkuperäisestä järjestelmästä irrotettuna, mikä helpottaa sisällön tulkitsemista valmiiden työkalujen avulla. Kirjallisuudesta ilmeni, että myös DBML-dokumenttien tarkastelu on teoriassa mahdollista XPath- ja XQuery-työkalujen avulla, mutta tämä ei ole käytännössä mielekästä DBML-dokumenttien suuren koon vuoksi.

FrepDB-järjestelmän prototyyppiin toteutettu vaihtoehtoinen tapa tarkastella tietokannan sisältöä perustuu sisäänrakennettuun ontologiaselaimeen. Ontologiaselaimen avulla tietokannan sisältöä voi selailla ontologialinkkejä seuraamalla, mikä voi auttaa hahmottamaan tietokannan rakennetta ja eri osien välisiä yhteyksiä.

XArch-järjestelmässä puolestaan on sisäänrakennettu toiminto, jonka avulla eri tietokantaotoksia voi selailla. Järjestelmää varten on kehitetty erityinen XAQL-kyselykieli, jonka avulla tietokantaan voidaan suorittaa kyselyjä. XAQL mahdollistaa tietokannan eri otosten tarkastelemisen, sekä tietokannan versiohistorian tarkastelemisen. XArch ainoana tutkituista säilytysmenetelmistä ei tarjonnut mahdollisuutta arkistoidun tietokannan palauttamiseksi tietokannanhallintajärjestelmään.

SIARD 2.0 -muodossa tallennettujen tietokantojen tarkastelua varten on kehitetty Database Visualization Toolkit -niminen työkalu (ks. luku 3.1.4). Sen avulla voidaan tarkastella muun muassa tietokannan rakennetta, kuvailutietoja ja tietokannan toiminnallisuuteen liittyviä tietoja, sekä visualisoida tietokannan taulujen välisiä relaatioita. Database Visualization Toolkit sisältää myös tiedonhakumahdollisuuden. Hakua on mahdollista rajata erilaisten kriteerien, kuten ajan, perusteella. Hakutulokset on mahdollista viedä PDF- tai CSV-tiedostoon myöhempää tarkastelua ja käsittelyä varten.

Säilytysmenetelmä	Haku- ja prosessointimahdollisuudet
DBML	XPath, XQuery
DBPreserve	XPath, XQuery
FrepDB	Ontologiaselain. Vain selailu, ei hakumahdollisuutta.
SIARD 1.0	SIARD Suite Viewer. Vain selailu, ei hakumahdollisuutta.
SIARD 2.0	Database Visualization Toolkit -työkalu. Fasetoitu hakumahdollisuus.
XArch	Sisäänrakennettu graafinen käyttöliittymä. Tiedon haku ja prosessointi XAQL-kyselykielelellä.

Taulukko 6: Tietokannan haku- ja prosessointimahdollisuudet eri säilytysmenetelmissä.

6 JOHTOPÄÄTÖKSET JA POHDINTAA

Tutkimuksessa tunnistettiin kaksi XML-konversioon pohjautuvaa säilytysformaattia, jotka ovat tällä hetkellä yleisessä käytössä: DBML ja SIARD. Näiden lisäksi tunnistettiin kaksi edellä mainittuihin säilytysformaatteihin perustuvaa säilytysmenetelmää, jotka olivat vielä prototyyppiasteella, DBPreserve ja FrepDB, sekä yksi pitkäaikaissäilytysjärjestelmä, XArch, joka mahdollistaa lukuisien tietokantaotosten liittämisen yhteen. Säilytysmenetelmiä tarkasteltiin E-ARK-projektin yhteydessä määriteltyjen keskeisten ominaisuuksien valossa (ks. luku 2.4). Tutkimuksesta selvisi, että säilytysmenetelmissä on puutteita erityisesti tietokannan käyttäytymiseen liittyvien ominaisuuksien säilymisen suhteen. Huolestuttavaa oli, että edes relaatiorakenteen säilyminen ei ollut kaikkien menetelmien kohdalla itsestään selvää.

XArchin tapauksessa kirjallisuudessa oli puutteita, eikä relaatiorakenteen säilymistä ollut erikseen käsitelty. Kirjallisuus viittasi kuitenkin siihen, että relaatiorakennetta ei säilytetä alkuperäisessä muodossa lainkaan, vaan tietokannan normalisaatio puretaan arkistointiprosessissa. Relaatiorakenteen säilyttäminen alkuperäisessä muodossa on relaatiotietokantojen säilytyksessä ensisijaisen tärkeää, sillä tietokannan relaatiot vaikuttavat siihen, kuinka tietokantaa käytetään, ja kuinka sen sisältöä tulkitaan ja ymmärretään (ks. Ferreira 2016, 19). Relaatiorakenteen muuttuminen vaikuttaa säilytetyn tietokannan autenttisuuteen ja luotettavuuteen, sekä sen luettavuuteen ja ymmärrettävyyteen. Puhuttaessa tietokannan autenttisuudesta ja luotettavuudesta Freitas (2012, 31) korosti tietokannan alkuperäisen merkityksen säilymistä sellaisena kuin se oli alun perin tarkoitettu (Freitas 2012, 31). Normalisaation purkaminen muuttaa tietokannan alkuperäistä merkitystä (ks. mm. Digital Preservation Testbed 2003, 21). Tämän valossa XArch-järjestelmää ei voi suositella relaatiotietokantojen ensisijaiseksi säilytysmenetelmäksi.

Tutkimuksen kannalta yllättävää oli, kuinka vähän tietokannan käyttäytymiseen liittyvien ominaisuuksien säilyttämistä alan tutkimuksessa oli painotettu. Relaatiotietokannan käyttäytymiseen liittyvien ominaisuuksien säilymisestä on tutkimuskirjallisuudessa puhuttu vasta viime aikoina, ja tutkimus osoitti, että ennen SIARD 2.0:n kehittämistä ei relaatiotietokantojen säilytysmenetelmien tutkimuksessa ole näitä ominaisuuksia otettu huomioon lainkaan. Ferreiran (2016, 19) mukaan tietokannan käyttäytymiseen liittyvät ominaisuudet, kuten herättimet, rajoittimet ja funktiot, kuuluvat relaatiotietokannan kes-

keisiin säilytettäviin ominaisuuksiin, sillä ainoastaan niiden avulla voidaan ymmärtää, kuinka tietokanta on aktiivikäytössä toiminut ja käyttäytynyt (Ferreira 2016, 19). Tämän vuoksi on suositeltavaa, että nämä seikat otetaan huomioon myös Suomessa niin julkishallinnon kuin yksityistenkin organisaatioiden relaatiotietokantojen säilytysratkaisuja valittaessa ja kehitettäessä. Säilytettävän tietokannan käyttäytymiseen liittyvien ominaisuuksien huolellisella dokumentoinnilla voidaan helpottaa organisaation toiminnasta kertovien tietokantojen tulkitsemista tulevaisuudessa.

Tutkimuksen parasta antia oli tarkastella, kuinka pääsy säilytettävien tietokantojen tietosisältöihin on eri säilytysmenetelmissä toteutettu. Suurimmassa osassa tarkasteltuja tutkimuksia painopiste oli kuitenkin säilytysmenetelmän teknisessä toteutuksessa. Tietokantojen haku- ja prosessointimahdollisuuksiin liittyvä kirjallisuuden osuus jäi hyvin pieneksi, mikä koettiin huomattavaksi puutteeksi. Tietokantojen haku- ja prosessointimahdollisuuksia tarkasteltaessa edukseen nousivat SIARD 2.0, sekä sille kehitetty työkalu, Database Visualization Toolkit (ks. Ferreira ym. 2016). Tiedonhaku helppokäyttöisellä, web-pohjaisella sovelluksella, sekä mahdollisuus viedä hakutulokset PDF- tai CSV-tiedostoon myöhempää tarkastelua varten, helpottavat datan hakemista ja käsittelyä, ja osoittavat suunnan tulevaisuuden pitkäaikaissäilytysjärjestelmien kehitykselle.

Myös DBPreserve-ohjelmiston prototyyppi osoittautui lupaavaksi säilytysmenetelmäksi. DBPreserven avulla tietovarastomuodossa arkistoitua tietokantaa voidaan analysoida ja käsitellä XML:n XQuery- ja XPath-työkalujen avulla. Tietovarastoksi muunnettu tietokanta on helpommin tulkittavissa ja analysoitavissa alkuperäisestä käyttökontekstista irrotettuna kuin normalisoitu relaatiotietokanta, mikä voidaan nähdä pitkäaikaissäilytyksen kannalta edulliseksi ominaisuudeksi. Koska DBPreserve säilyttää tietokannasta myös alkuperäisen relaatiorakenteen mukaisen version SIARD-muodossa, eivät tietokannan autenttisuus, luotettavuus ja ymmärrettävyys vaarannu arkistointiprosessissa. DBPreserven eduiksi voidaan katsoa myös se, että XQuery-kyselykieli on SQL-kyselykieltä yksinkertaisempi, mikä voi helpottaa järjestelmän käyttämistä. XQueryn avulla on lisäksi mahdollista toteuttaa valmiita, organisaation tarpeisiin räätälöityjä työkaluja säilytetyn tietokannan tarkastelua, käsittelyä ja analysointia varten. Tulevaisuus näyttää, mihin DBPreserven kehitys johtaa, ja tuleeko ohjelmisto avoimeen levitykseen.

Toinen tutkimuksessa huomiota herättänyt säilytysratkaisu oli FrepDB-järjestelmän prototyyppi, jossa tietokannan rakennetta oli pyritty yksinkertaistamaan ja pääsyä tietosisäl-

töön helpottamaan ontologioiden avulla. Tietokannasta muodostettiin OWL-ontologiakielen avulla käsitteellinen malli, joka liitettiin tietokannan säilytystiedostoon. Ontologioiden vahvuutena on, että ne mahdollistavat tietokannan sisällön ja rakenteen tulkinnan koneellisesti, mikä mahdollistaa eri järjestelmien välisen tiedonvaihdon ja yhteistoiminnallisuuden, sekä edesauttaa automaattisten järjestelmien toteuttamista. Koska projekti oli tutkimusta tehtäessä prototyyppivaiheessa, ja kirjallisuutta oli saatavilla vähän, on mahdoton tehdä arvioita säilytysmenetelmän toiminnallisuudesta tai käytännöllisyydestä. Sekä DBPreserve että FrepDB osoittavat kuitenkin innovatiivisuutta, jota tietokantojen pitkäaikaissäilytyksen tutkimuksessa tarvitaan. Näiden kaltaisia säilytysratkaisuja tulisi tutkia ja kehittää lisää, jotta arkistoitavien tietokantojen sisältö voidaan saattaa helpommin käytettäviksi ja saavutettaviksi, ja arkistoitujen tietokantojen koko potentiaali tutkimuskäytölle saadaan hyödynnettyä.

Eräs tietokantojen pitkäaikaissäilytyksen nykymenetelmiin liittyvä ongelma on se, että tietokannat tai niiden otokset täytyy ladata tarkasteltavaksi yksi kerrallaan, mikä hankaloittaa ajallisten kyselyiden ja pitkittäistutkimusten tekemistä (ks. Delve ym. 2014). Tämän vuoksi on erityisen tärkeää keskittyä kehittämään sellaisia tietokantojen käsittelyä helpottavia tekniikoita, joiden avulla tietoa voidaan hakea useista tietokannoista samanaikaisesti. Tutkimuksessa tarkasteltu XArch-säilytysjärjestelmä ratkaisee osan näistä ongelmista. XArch-järjestelmässä kaikki tietokannan otokset voidaan liittää samaan säilytystiedostoon, mikä mahdollistaa ajallisten kyselyiden tekemisen ja tietokannan muutoshistorian tarkastelemisen. XArchin heikkoutena on kuitenkin yksinkertainen, hierarkkinen XML-muoto. Tutkimuskirjallisuudesta tehdyn tulkinnan perusteella XArch säilyttää tietokannan relaatiot ja muut keskeiset ominaisuudet heikosti. XArch ei myöskään mukaudu OAIS-viitemallin mukaisessa pitkäaikaissäilytysjärjestelmässä käytettäväksi, minkä vuoksi sitä ei voi suositella ainoaksi ratkaisuksi organisaation tehtävien hoidossa syntyvien tietokantojen säilytykseen. Se voi kuitenkin toimia oivallisena oheismenetelmänä tietokantojen versiohistorian säilyttämiseksi, kunhan nämä rajoitukset otetaan huomioon, ja järjestelmää käytetään jonkin autenttisuuden ja luotettavuuden takaavan säilytysmenetelmän ohella. Puutteistaan huolimatta XArch toimii suunnannäyttäjänä relaatiotietokantojen pitkäaikaissäilytyksen tutkimukselle. Alan tulevassa tutkimuksessa ja ohjelmistokehityksessä on hyvä ottaa huomioon mahdollisuus useiden tietokantaotosten yhdistämiseksi helposti vertailtavaan, ajalliset kyselyt mahdollistavaan muotoon. XArch toimii oivallisena esimerkkinä ja inspiraationa sille, millaisia ominaisuuksia tietokantojen pitkäaikaissäilytysjärjestelmiltä voidaan tulevaisuudessa odottaa.

Kaiken kaikkiaan tutkielman tulokset kertovat, että relaatiotietokantojen pitkäaikaissäilytyksen tutkimus on vielä varhaisessa vaiheessa. Lisää tutkimusta tarvitaan erityisesti helppokäyttöisistä ja käyttäjäystävällisistä ratkaisuista, joiden avulla tarjota pääsy arkistoitujen tietokantojen tietosisältöihin. Tutkimuskäytössä on tarvetta erityisesti sellaisille säilytysratkaisuille ja hakutyökaluille, jotka mahdollistavat useita eri tietokantaotoksia kattavat, ajalliset kyselyt. Tietovarastoteknologian hyödyntäminen relaatiotietokantojen säilytyksessä on mielenkiintoinen tutkimusaihe, sillä tietovarastoteknologian avulla voidaan potentiaalisesti helpottaa ja tehostaa säilytettävien tietokantojen tarkastelua ja analysointia. Eräs keskeinen jatkotutkimuksen kohde on tietokantasovellusten säilyttäminen, jota ei ollut tämän tutkielman puitteissa mahdollista käsitellä lainkaan.

Kirjallisuuden pohjalta keskeiseksi relaatiotietokantojen säilytysmenetelmäksi nousi SIARD 2.0 -säilytysformaatti, sekä sille toteutettu Database Visualization Toolkit -työkalu. SIARD 2.0 osoittautui tutkimuksessa tarkastelluista säilytysformaateista monipuolisimmaksi ja nykyaikaisimmaksi ratkaisuksi tietokantojen säilytykseen. Tutkimuksen perusteella SIARD 2.0 vaikuttaa olevan vakiintumassa relaatiotietokantojen *de facto* -säilytysstandardiksi. Tähän viittaa muun muassa se, että monissa kansainvälisissä pitkäaikaissäilytysprojekteissa, kuten MIXED, RODA ja E-ARK, päädyttiin valitsemaan SIARD ensisijaiseksi relaatiotietokantojen säilytysformaatiksi. SIARD on myös laajalti käytetty formaatti kansainvälisellä tasolla; se on käytössä jo 59 eri maassa (Swiss Federal Archives 2016).

Tutkimuksen valossa voidaan todeta, että SIARD 2.0 on tällä hetkellä monipuolisin, joustavin ja työkaluiltaan kehittynein vaihtoehto relaatiotietokantojen pitkäaikaissäilytysformaatiksi, ja sitä voidaan suositella yritysten ja julkishallinnon organisaatioiden tarpeisiin myös Suomessa. Tietokantojen pitkäaikaissäilytys on kuitenkin digitaalisen pitkäaikaissäilytyksen kenttä, johon liittyy vielä paljon ratkaisemattomia ongelmia, ja jatkotutkimukselle on voimakas tarve, jotta ainutlaatuista tietoa sisältävien ja arvoltaan korvaamattomien tietokantojemme säilyminen autenttisena, käytettävänä ja saavutettavana voidaan taata nyt ja tulevaisuudessa.

LÄHTEET

Aldeias, C., David, G. ja Ribeiro, C., 2011. DWXML - A Preservation Format for Data Warehouses! Teoksessa A. Simões (toim.) *Ninth National Conference on XML, Associated Technologies and Applications (XATA 2011)*. Vila do Conde, s. 115–126.

Aldeias, C. F. P., 2011. Open Archival Information Systems for Database Preservation. FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO. Saatavilla osoitteessa: https://repositorio-aberto.up.pt/bitstream/10216/63425/1/000150055.pdf [Viitattu 16. syyskuuta, 2016].

Arkistolaitos, 2009. SÄHKE2-määräys. Liite 1. Asiakirjallisten tietojen metatietojen tuottamisen periaatteet.

Arkistolaki 1994/831. Saatavilla osoitteessa: http://www.finlex.fi/fi/laki/ajantasa/1994/19940831.

Ashley, K., 2004. The preservation of databases. *VINE: The Journal of Information and Knowledge Management Systems*, 34(2), s. 66–70. Saatavilla osoitteessa: http://www.emeraldinsight.com/doi/abs/10.1108/03055720410551075 [Viitattu 16. syyskuuta, 2016].

Atos, 2014. Digital Preservation in the Age of Cloud and Big Data. Saatavilla osoitteessa: https://atos.net/content/dam/global/ascent-whitepapers/ascent-whitepaper-digital-preservation-in-the-age-of-cloud-and-big-data.pdf [Viitattu 10. tammikuuta, 2017].

Aveyard, H., 2010. Doing A Literature Review In Health And Social Care: A Practical Guide. McGraw-Hill Education.

Avoin tiede ja tutkimus, 2017. Tutkimusaineistojen tiedostomuodot ja pitkäaikaissäilytyskelpoisuus. Selvityksen loppuraportti. Saatavilla osoitteessa: http://avointiede.fi/documents/10864/12232/Tutkimusaineistojen_tiedostomuodot_loppuraportti.pdf/24557e81-f504-4383-9a27-304e09b27e94 [Viitattu 24. maaliskuuta, 2017].

Becker, C., Rauber, A., Heydegger, V., Schnasse J. ja Thaller, M., 2008. Systematic Characterisation of Objects in Digital Preservation: The eXtensible Characterisation Languages. *Journal of Universal Computer Science*, 14(18), s. 2936–2952.

Benedikt, M., Chan, C., Fan, W., Rastogi, R., Zheng, S. ja Zhou, A., 2002. DTD-Directed Publishing with Attribute Translation Grammars. Teoksessa *Proceedings of the 28th International Conference on Very Large Data Bases (VLDB)*. s. 838–849. Saatavilla osoitteessa: http://homepages.inf.ed.ac.uk/wenfei/papers/vldb02-atg.pdf [Viitattu 16. huhtikuuta, 2017].

Das Bundesarchiv, 2011. Das Digitale Archiv des Bundesarchivs. Saatavilla osoitteessa: https://www.bundesarchiv.de/imperia/md/content/bundesarchiv_de/fachinformation/informationstechnologie/digitalesarchiv/brosch_re_das_digitale_archiv_des_bundesarchivs_stand_august_2011.pdf [Viitattu 24. maaliskuuta, 2017].

Choo, C.W., Detlor, B. ja Turnbull, D., 2000. Information seeking on the Web: An integrated model of browsing and searching. *First Monday*, 5(2). Saatavilla osoitteessa: http://journals.uic.edu/ojs/index.php/fm/article/view/729/638.

CINES, 2017. FACILE - Service de validation de formats. Version du validateur de formats: 3.4.5. Saatavilla osoitteessa: https://facile.cines.fr/ [Viitattu 24. maaliskuuta, 2017].

Codd, E. F., 1970. A Relational Model of Data Large Shared Data Banks. *Information Retrieval*, 13(6), s. 377–387.

Connolly, T. ja Begg, C., 2015. Database Systems. A Practical Approach to Design, Implementation and Management. 6. painos. Pearson.

Conway, P., 2010. Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. *Library Quarterly*, 80(1), s. 61–79.

Dappert, A. ja Enders, M., 2010. Digital Preservation Metadata Standards. *Information Standards Quarterly*, 22(2), s. 4–13.

Data Archiving and Networked Services (DANS), 2010. Standard Data Formats for Preservation (SDFP). Saatavilla osoitteessa: https://sites.google.com/a/datanetworkservice.nl/mixed/documentation [Viitattu 12. huhtikuuta, 2017].

Data Archiving and Networked Services (DANS), 2015. Preferred formats. Saatavilla osoitteessa: https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf [Viitattu 12. huhtikuuta, 2017].

Dell Software, 2015. Dell Survey: Structured Data Remains Focal Point Despite Rapidly Changing Information Management. Saatavilla osoitteessa: http://www.dell.com/learn/us/en/ph/press-releases/2015-04-15-dell-survey [Viitattu 8. lokakuuta, 2015].

Delve, J., Schmidt, R. ja Aas, K., 2014. LONG-TERM PRESERVATION OF DATABASES THE MEANINGFUL WAY. Teoksessa *DLM Forum Triennial Conference 2014, November 10-14. Lisbon*.

Digital Preservation Testbed, 2003. From Digital Volatility to Digital Permanence: Preserving databases. Technical Report, Dutch National Archives. Saatavilla osoitteessa: http://www.nationaalarchief.nl/sites/default/files/docs/kennisbank/volatility-permanence-databases-en.pdf [Viitattu 8. lokakuuta, 2016].

eCH-0165, 2016. SIARD format specification. Version 2.0. eCH e-Government Standards. Saatavilla osoitteessa: https://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0165&documentVersion=2.0 [Viitattu 24. maaliskuuta, 2017].

Elmasri, R. ja Navathe, S., 1994. Fundamentals of Database Systems. 2. painos. The Benjamin/Cummings Publishing Company, Inc.

van Essen, M., de Rooij, M., Roberts, B. ja van den Dobbelsteen, M., 2011. Case Study: Database Preservation at the National Archives of the Netherlands. Planets project report. Saatavilla osoitteessa: http://openpreservation.org/system/files/Database%20archiving%20review.pdf [Viitattu 25. huhtikuuta, 2017].

Factor, M., Henis, E., Naor, D., Rabinovici-Cohen, S., Reshef, P., Ronen, S., Michetti, G. ja Guercio, M., 2009. Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage. Teoksessa *Proceedings of the 1st Workshop on the Theory and Practice of Provenance (TAPP'09). February 23*. San Francisco, CA. Saatavilla osoitteessa: https://www.usenix.org/legacy/event/tapp09/tech/ [Viitattu 17. syyskuuta, 2016].

Faniel, I. M. ja Yakel, E., 2011. Significant properties as contextual metadata. *Journal of Library Metadata*, 11(3–4), s. 155–165.

Ferreira, B., 2016. Database Preservation Toolkit. A relational database conversion and normalization tool. Universidade do Minho. Saatavilla osoitteessa: http://hdl.handle.net/1822/43479 [Viitattu 25. huhtikuuta, 2017].

Ferreira, B., Faria, L., Ramalho, J. ja Ferreira, M., 2016. Database Preservation Toolkit. A relational database conversion and normalization tool. Teoksessa *Proceedings of the 13th International Conference on Digital Preservation (iPRES2016). October 3-6, 2016*. Bern, Sveitsi. Saatavilla osoitteessa: https://ipr16.organizers-congress.org/frontend/organizers/media/iPRES2016/_PDF/IPR16.Proceedings_4_Web_Broschuere_Link.pdf [Viitattu 25. huhtikuuta, 2017].

Fink, A., 2013. Conducting Research Literature Reviews: From the Internet to Paper. 4. painos. SAGE Publishing.

Freitas, R., 2012. Relational Databases Digital Preservation. Universidade do Minho, Portugali. Saatavilla osoitteessa: http://repositorium.sdum.uminho.pt/bitstream/1822/25655/1/Ricardo André Pereira Freitas.pdf [Viitattu 10. tammikuuta, 2017].

Grönfors, M., 2011. Laadullisen tutkimuksen kenttätyömenetelmät. Toim. H. Vilkka. Hämeenlinna: SoFia-Sosiologi-Filosofiapu Vilkka.

Hakala, J., 2002. Elektronisten aineistojen säilyttämisestä. Teoksessa Arkisto. Arkistoyhdistyksen Julkaisuja, 8. s. 19–38.

Hakala, J., 2014. Pitkäaikaissäilytyksen standardit. *Tietolinja - Kansalliskirjaston verkkolehti*. Saatavilla osoitteessa: http://urn.fi/URN:NBN:fi-fe2014120552176 [Viitattu 25. huhtikuuta, 2017].

Hedstrom, M., 1998. Digital Preservation: A Time Bomb for Digital Libraries. *Computers and the Humanities*, 31, s. 189–202.

Henttonen, P., 1999. Atk-aineistojen arkistoinnin haasteita. Teoksessa *Arkisto. Arkistoyhdistyksen julkaisuja 6*. s. 23–64.

Heuscher, S., Järmann, S., Keller-Marxer, P. ja Möhle, F., 2004. Providing Authentic Long-term Archival Access to Complex Relational Data. Teoksessa *European Space Agency Symposium. "Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data". 5-7 October*. Frascati, Italia. Saatavilla osoitteessa: https://arxiv.org/abs/cs/0408054 [Viitattu 10. tammikuuta, 2017].

Hirsjärvi, S., Remes, P. ja Sajavaara, P., 2007. Tutki ja kirjoita. 13. painos. Tammi.

van Horik, R. ja Roorda, D., 2011. Migration to Intermediate XML for Electronic Data (MIXED): Repository of Durable File Format Conversions. *The International Journal of Digital Curation*, 2(6).

van Horik, R. ja Roorda, D., 2009. MIXED: Repository of Durable File Format Conversions. Teoksessa *Proceedings of the Sixth International Conference on the Preservation of Digital Objects (iPRES2009), California Digital Library, UC Office of the President*. s. 194–197. Saatavilla osoitteessa: http://escholarship.org/uc/item/8h39210x [Viitattu 7. maaliskuuta, 2017].

ISO/IEC 9075-1:2016 - Information technology -- Database languages -- SQL -- Part 1: Framework (SQL/Framework). *Standards Catalogue*. Saatavilla osoitteessa: https://www.iso.org/standard/63555.html [Viitattu 5. tammikuuta, 2017].

ISO 9075:1987 - Information processing systems -- Database language -- SQL. *Standards Catalogue*. Saatavilla osoitteessa: https://www.iso.org/standard/16661.html [Viitattu 5. tammikuuta, 2017].

JUHTA, 2004. JHS 143 -suositus. Asiakirjojen kuvailun ja hallinnan metatiedot. Julkisen hallinnon tietohallinnon neuvottelukunta JUHTA. Saatavilla osoitteessa: http://docs.jhs-suositukset.fi/jhs-suositukset/JHS143/JHS143.pdf [Viitattu 24. maaliskuuta, 2017].

JUHTA, 2017. Rekisteritiedon metatiedot. Hankesuunnitelma. Julkisen hallinnon tietohallinnon neuvottelukunta JUHTA. Saatavilla osoitteessa: http://jhs-suositukset.netum.fi/c/document_library/get_file?uuid=bd4954b6-de25-4b0a-abf2-40a303908ad4&groupId=14 [Viitattu 24. maaliskuuta, 2017].

Lawrence, A., 2001. New Perspectives On Preserving Documents. *National Underwriter Property & Casualty/Risk & Benefits Management Edition*, 105(23).

Lee, K-H., Slattery, O., Lu, R., Tang, X. ja McCrary, V., 2002. The State of the Art and Practice in Digital Preservation. *Journal of Research of the National Institute of Standards and Technology*, 107(1), s. 93–106.

Lin, L., Ramaiah, C. ja Wal, P., 2003. Problems in the preservation of electronic records. *Library Review*, 52(3), s. 117–125. Saatavilla osoitteessa: http://dx.doi.org/10.1108/00242530310465924 [Viitattu 10. tammikuuta, 2017].

Lybeck, J., 2006. ARKISTOT YHTEISKUNNAN TOIMIVA MUISTI. Asiakirjahallinnon ja arkistotoimen oppikirja.

Müller, H., 2009. Database archiving, *DCC Briefing Papers: Introduction to Curation*. Edinburgh: Digital Curation Centre. Saatavilla osoitteessa:

http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/databasearchiving [Viitattu 16. huhtikuuta, 2017].

Müller, H., Buneman, P. ja Koltsidas, I., 2008. XArch: Archiving Scientific and Reference Data. Teoksessa *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. Vancouver: ACM New York, s. 1295–1298.

Pulkkinen, M. P., 1994. Relaatiotietokantojen arkistointi. Teoksessa *Arkisto. Arkistoyhdistyksen julkaisuja 5*. s. 51–66.

Rahman, A. U., Muzammal, M., David, G. ja Ribeiro, C., 2015. Database Preservation: The DBPreserve Approach. *International Journal of Advanced Computer Science and Applications*, 6(12).

Rahman, A. U., David, G. ja Ribeiro, C., 2012. SIARD Archive Browser. Teoksessa Zaphiris, P., Buchanan, G., Rasmussen, E. ja Loizides, F. (toim.) *Theory and Practice of Digital Libraries. TPDL 2012. Lecture Notes in Computer Science, vol 7489*. Springer, Berlin, Heidelberg, s. 496–499.

Ramalho, J. C., Faria, L., Silva, H. ja Coutada, M., 2014. Database Preservation Toolkit: a flexible tool to normalize and give access to databases. Teoksessa *DLM Forum - 7th triennial conference. Making the Information Governance Landscape in Europe. 10.-14. November 2014. Lissabon, Portugal*. Biblioteca Nacional de Portugal. Saatavilla osoitteessa: http://hdl.handle.net/1822/35183 [Viitattu 25. huhtikuuta, 2017].

Ramalho, J. C., Ferreira, M., Faria, L., Castro, R., Barbedo, F. ja Corujo, L., 2008. RODA and Crib A Service-Oriented Digital Repository. Teoksessa *Proceedings of the Fifth International Conference on Preservation of Digital Objects - iPRES 2008. The British Library, London. 29-30 September*. s. 235-241. Saatavilla osoitteessa: https://www.bl.uk/ipres2008/ipres2008-proceedings.pdf [Viitattu 13. lokakuuta, 2016].

RODA, 2016. RODA - Repository of Authentic Digital Objects. KEEP SOLUTIONS. Saatavilla osoitteessa: http://www.keep.pt/en/produtos/roda/ [Viitattu 13. lokakuuta, 2016].

Saldana, J., Leavy, P. ja Beretvas, N., 2014. Fundamentals of Qualitative Research. Oxford University Press.

Salminen, A., 2011. Mikä kirjallisuuskatsaus? Johdatus kirjallisuuskatsauksen tyyppeihin ja hallintotieteellisiin sovelluksiin. Teoksessa *Vaasan yliopiston julkaisuja. Opetusjulkaisuja 62. Julkisjohtaminen 4. Vaasa*.

Shepherd, E. ja Smith, C., 2000. The Application of ISAD(G) to the Description of Archival Datasets. *Journal of the Society of Archivists*, 21(1), s. 55–86. Saatavilla osoitteessa: http://www.tandfonline.com/action/journalInformation?journalCode=cjsa20 [Viitattu 10. tammikuuta, 2017].

Swiss Federal Archives, 2001. Archiving of Electronic Digital Data and Records in the Swiss Federal Archives (ARELDA). Saatavilla osoitteessa: https://web.archi-

ve.org/web/20051220231227/http://www.bar.admin.ch/webserverstatic/docs/e/arelda_expose_0301_e.pdf [Viitattu 17. kesäkuuta, 2017].

Swiss Federal Archives, 2010. Save Your Databases - The SIARD Relational Database Archiving Solution. Factsheet. Saatavilla osoitteessa: https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html [Viitattu 25. kesäkuuta, 2017].

Swiss Federal Archives, 2016. Archiving of Databases: SIARD Suite. Distribution of SIARD. Saatavilla osoitteessa: https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html [Viitattu 25. kesäkuuta, 2017].

W3C, 2008. Extensible Markup Language (XML) 1.0 (Fifth Edition). W3C Recommendation. Saatavilla osoitteessa: https://www.w3.org/TR/xml/ [Viitattu 29. toukokuuta, 2017].

W3C, 2016. XML Path Language (XPath). Version 1.0. W3C Recommendation. Saatavilla osoitteessa: https://www.w3.org/TR/xpath/ [Viitattu 29. toukokuuta, 2017].

W3C, 2017. XQuery 3.1: An XML Query Language. W3C Recommendation. Saatavilla osoitteessa: https://www.w3.org/TR/xquery-31/ [Viitattu 29. toukokuuta, 2017].

Wilson, A., 2008. Significant Properties of Digital Objects. JISC Significant Properties Workshop, British Library, 7. huhtikuuta, 2008. Saatavilla osoitteessa: https://pdfs.semanticscholar.org/4fb6/0523dd6c9c8be19d79d8fc8bedb0ad75c474.pdf [Viitattu 25. kesäkuuta, 2017].